Technical Report
1225

Estimating Classifier Accuracy Using Noisy Expert Labels

J.T. Holodnak
J.T. Matterer
W.W. Streilein

31 January 2018

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

© 2017 MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

Massachusetts Institute of Technology
Lincoln Laboratory

Estimating Classifier Accuracy Using Noisy Expert Labels

J.T. Holodnak
J.T. Matterer
W.W. Streilein

Group 58

Technical Report 1225

31 January 2018

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

In this work, we present an empirical comparison of statistical methods that estimate the accuracy of a classifier using noisy expert labels. We are motivated by the application of machine learning to difficult problems for which even experts may be unable to provide an authoritative label for every data instance. Several estimators have been recently proposed in the literature, but prior empirical work to evaluate the applicability of these estimators to real-world problems is limited. We apply the estimators to labels simulated from three models of the expert labeling process and also four real datasets labeled by human experts. Our simulations reveal the importance of the accuracy of the classifier relative to the experts and confirm that conditional dependence between experts negatively impacts estimator performance. On two of the real datasets, the estimators clearly outperformed the baseline majority vote estimator, supporting their use in applications. We also briefly examine the utility, in terms of increasing or decreasing confidence in an estimator's output, of a few diagnostics that can be applied to the expert labels.

This page intentionally left blank.

ACKNOWLEDGMENTS

Thanks to David Weller-Fahy, Seth Webster, Charles Dagli, and Elizabeth Godoy for help with data.

This page intentionally left blank.

TABLE OF CONTENTS

		Page
Abstract		iii
Acknowledgments		v
List of Figures		ix
List of Tables		xiii
1.	INTRODUCTION	1
1.1	Prior empirical work	1
1.2	Our focus areas	2
1.3	Organization	2
2.	NOTATION AND PROBLEM SETTING	3
3.	ACCURACY ESTIMATORS	5
3.1	Baselines	5
3.2	Agreement-based estimators	5
3.3	Graphical Models (MLE and BEE)	7
3.4	Covariance-based Estimator (COV)	7
4.	SIMULATIONS	9
4.1	Models considered	9
4.2	Model parameterization	10
4.3	Results	11
4.4	Summary	19
5.	EXPERIMENTS	21
5.1	Job Category	21
5.2	Cyber Social Media Conversations I	25
5.3	Cyber Social Media Conversations II	28
5.4	Fake News	31
5.5	Summary	35
6.	DIAGNOSTIC TESTS	37
6.1	Inter-rater reliability	37

TABLE OF CONTENTS
(Continued)

Page

6.2 Detecting dependence between experts 37

7. CONCLUSION AND FUTURE WORK 43

7.1 Future Work 44

References 45

LIST OF FIGURES

Figure No.		Page
1	Graphical model of label generation.	7
2	Models of label generation.	9
3	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots. The actual accuracies are 0.8 times the noisy accuracies.	12
4	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the copy model of dependence for each estimator. The accuracies of the experts and classifier are given in the title of the individual plots.	14
5	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the difficulty model of dependence for each estimator. The accuracy penalty on hard instances and the accuracy of the classifier on easy instances are given in the title of the individual plots. The overall accuracy of the experts or classifier is $\frac{x}{2} + \frac{x(1-y)}{2}$, where x is the accuracy on easy instances and y is the hard instance penalty.	15
6	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the copy model of dependence for each estimator. The value of ρ and the accuracy of the classifier are given in the title of the individual plots.	16
7	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence with one group for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots. The actual accuracies are 0.8 times the noisy accuracies.	17
8	Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence with three groups for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots.	18
9	Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.	22

LIST OF FIGURES
(Continued)

Figure No.		Page
10	Signed errors for 10,000 resamples of the Job Category dataset for each estimator.	23
11	Absolute errors of MLE (left) and AGR (right) for all possible subsets of between two and seven experts, on the Job Category dataset.	24
12	Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.	26
13	Signed errors for 10,000 resamples of the Social Media I dataset for each estimator.	26
14	Absolute errors of MLE (left) and AGR (right) for all possible subsets of between two and six experts, on the Social Media I dataset.	27
15	Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.	29
16	Signed errors for 10,000 resamples of the Social Media II dataset, for each estimator.	29
17	Absolute errors of MLE (left) and AGR (right) for all possible subsets of between two and six experts, on the Social Media II dataset.	30
18	Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.	32
19	Signed errors for 10,000 resamples of the Fake News dataset, for each estimator.	33
20	Absolute errors of MLE (left) and AGR (right) for all possible subsets of between two and six experts, on the Fake News dataset.	34
21	Clusters returned by spectral clustering (highlighted in red for clarity in the bottom right) for 10,000 and 200 samples, when two and four clusters are requested. The correct number of clusters is two.	40
22	Clusters returned by spectral clustering (highlighted in red for clarity), when two, three, and four clusters are requested, using the matrix D for the job category experiment.	40
23	Clusters returned by spectral clustering, when two, three, and four clusters are requested, using the matrix of pairwise Yule's Q statistic for the job category experiment.	41

LIST OF FIGURES
(Continued)

Figure No. **Page**

24 Left: Error of MLE when applied to experts that come from only the classifier group, only the other group, or to experts from both groups. Right: Error of AGR for the same experiment. 41

This page intentionally left blank.

LIST OF TABLES

Table No.		Page
1	Mean absolute errors of the estimators across 10,000 resamples of the Job Category dataset.	23
2	Mean absolute errors of the estimators across 10,000 resamples of the Social Media I dataset.	27
3	Mean absolute errors of the estimators across 10,000 resamples of the Social Media II dataset.	30
4	Description of the classes in the original Buzzfeed dataset and which classes we used in our "Fake" and "Not Fake" classes.	31
5	Mean absolute errors of the estimators across 10,000 resamples of the Fake News dataset.	33
6	Errors for estimating the accuracy of the classifier, using all samples. The smallest error for each experiment is in bold.	36
7	Fleiss' κ for experts from the experiments, as well as simulated experts from the Copy and Group models.	38
8	Proportion of times that spectral clustering returns the correct group structure for increasing numbers (k) of groups.	39

This page intentionally left blank.

1. INTRODUCTION

Estimating the accuracy of a classifier is usually a straightforward task: simply compare the classifier's predicted labels with the true labels for some test set. In applications where ground truth is not readily available, expert judgment is often used to determine the true labels. When these expert labels contain significant *label noise*, however, accuracy estimates derived from them may be unreliable. The common use of crowd workers (who may have limited expertise or dedication to the task) from web-based services such as Amazon Mechanical Turk for labeling data and the application of machine learning to problems difficult even for true human experts make the importance of accounting for label noise in evaluation clear.

In this paper, we consider the following setting: Suppose we have a dataset $\mathcal{D}$ of samples from $\mathcal{X}$ and assume each datapoint in $\mathcal{D}$ receives a label from a classifier and each of E experts. Our goal is to use these labels to estimate the accuracy of the classifier on $\mathcal{X}$. Training the classifier is a separate issue; it may be unsupervised or trained on a different labeled dataset.

Several approaches to estimating accuracy in this setting have been proposed, most in the last few years. Dawid and Skene [1] appear to have been one of the first to propose a non-trivial algorithm to estimate accuracy using noisy labels. More recently, several estimators have appeared in the literature [2–6].

In this paper, our primary goal is to empirically analyze the effectiveness of these estimators on both simulated and real data. In particular, we are interested in the case where the classifier is better than the human experts. This runs counter to most work in evaluation of systems; typically, one assumes that if there are errors in the ground truth (our expert labels), they are insignificant compared to the errors made by the system. As classifiers become increasingly sophisticated (and as more labeling tasks are performed via crowdsourcing) this assumption is becoming increasingly less valid. Some authors have noted problems with this assumption in the forecast evaluation literature as well [7, 8].

1.1 PRIOR EMPIRICAL WORK

As we intend our study to be primarily empirical, we now briefly describe how the estimators mentioned above have been empirically evaluated in prior work. Donmez et al. [2] present the most comprehensive experimental treatment, examining the robustness of their estimator to experts with a mix of accuracies on synthetic datasets, as well as real datasets with both humans and trained classifiers in the role of the "experts." While they note that experts are likely dependent in several of their datasets, they do not confirm this nor investigate the strength of dependence. Platanios et al. [3,6] apply their estimators to a natural language processing (NLP) dataset as well as a functional magnetic resonance imaging dataset. They compute a measure of *unconditional* dependence for the NLP data, but do not explore the effect of dependence in general. In addition, they do not perform experiments where data is gathered from human experts. Lehner [5] performs several experiments on simulated data, but all experiments appear to have conditionally independent experts. The author also considers the bias of the estimator and its relation to the accuracy of the classifier.

Finally, Jaffe et al. [4] apply their estimator only as a component of an ensemble method for inferring the true label.

1.2 OUR FOCUS AREAS

We note three major gaps in prior empirical evaluations. First, almost no prior work has applied these estimators to data labeled by human experts. They have mostly been applied to simulated data or to large datasets labeled by automated classifiers. Second, there is little to no understanding of how well these estimators perform when applied to labels from experts that are conditionally dependent. Third, prior work has not explicitly considered how the performance of the classifier relative to the experts impacts estimator error. In this paper, we address each of these issues through an extensive set of simulations and four experiments involving human experts.

1.3 ORGANIZATION

We now outline the remainder of this paper. In Section 2, we set our notation and discuss our problem setting. We describe several estimators from the literature and two baseline accuracy estimators in Section 3. Then, in Sections 4 and 5, we apply these methods to both simulated and real data. We describe one new model for simulating data and use two others from the literature in Section 4 to make inferences about the performance of the estimators as accuracy and expert dependence are varied. We show results for four experiments involving human experts in Section 5. Then, in Section 6, we discuss the extent to which diagnostics can be applied to a dataset labeled by experts to increase or decrease the confidence of a practitioner in the output of the method. Finally, we conclude and discuss opportunities for future work in Section 7.

2. NOTATION AND PROBLEM SETTING

Suppose we have a dataset $\mathcal{D} = \{x_n\}_{n=1}^N$ of samples from $\mathcal{X}$ with true labels $y_n \in \mathcal{Y} = \{1, \ldots, L\}$ that are unknown. The true labels occur according to the class prior distribution $\pi = (\pi_1, \ldots, \pi_L)$. Given a classifier $f^0 : \mathcal{X} \to \mathcal{Y}$, we denote its label on data point x_n as $\hat{y}_n^0 := f^0(x_n)$. Denote the classifier's confusion matrix entries as $\psi_{ll'}^0 := P(\hat{y}^0 = l' \,|\, y = l)$ for $1 \leq l \leq L,\ 1 \leq l' \leq L$.

Our goal is to estimate the classifier's accuracy $\alpha^0 = \sum_{l=1}^L \pi_l \psi_{ll}^0$. To do this, suppose we also have access to E experts $f^e : \mathcal{X} \to \mathcal{Y},\ 1 \leq e \leq E$ and that each expert labels all the points in our dataset. We denote the label of expert e on datapoint x_n as $\hat{y}_n^e := f^e(x_n)$. For each expert e, we denote the expert confusion matrix entries as $\psi_{ll'}^e := P(\hat{y}^e = l' \,|\, y = l)$ for $1 \leq l \leq L,\ 1 \leq l' \leq L$.

The accuracy estimators described previously consume labels from both the classifier and experts $\{(\hat{y}_n^0, \hat{y}_n^1, \ldots, \hat{y}_n^E)\}_{n=1}^N$ and output an estimate $\hat{\alpha}^0$ to α^0. We are interested in both the signed and absolute error between the true classifier accuracy and the estimated classifier accuracy, that is $\hat{\alpha}^0 - \alpha^0$ and $|\hat{\alpha}^0 - \alpha^0|$.

We note that all past work, with the exception of [5], has focused on the mean absolute error over the classifier and all the experts. Of most interest to us is the use case in which the experts are specifically used to estimate the accuracy of the classifier. As a result, we compute errors with respect to the performance of the classifier only, as described above.

Since our focus is to consider the application of these methods to studies involving labels gathered from human experts, we will consider the performance of these methods when the number of samples is small, that is $N \leq 200$. As a result, we will also consider a small number of classes. While accuracy does not depend on the number of classes, we find it inconsistent to consider a large number of classes and a small number of samples. In addition to the seeming inconsistency, some of the estimators compute per-class accuracy and as a result, will likely perform very poorly if applied to a many-class problem using a small number of samples. As a result, we will consider between two and five classes in this work. Previous work, except that by Lehner [5], has considered only two classes.

This page intentionally left blank.

3. ACCURACY ESTIMATORS

In this section, we describe a simple baseline accuracy estimator and five statistical accuracy estimators.

3.1 BASELINES

Majority Vote (MV) Perhaps the most natural way to use the expert labels to evaluate the classifier is to simply combine the expert labels via majority vote and regard this as the ground truth. Then, estimate accuracy as the percent agreement between the classifier labels and the labels derived from the expert majority vote.

3.2 AGREEMENT-BASED ESTIMATORS

The similar, but distinct, estimators below both make use of agreement between experts to estimate classifier accuracy.

3.2.1 Optimization based on pair-wise agreement rates (AGR-OPT)

Platanios et al. [3] use agreement rates between a set of classifiers/experts to estimate the accuracy of each. The core observation is that it is possible to express the probability of pairwise agreement, which is easily measured, as a function of accuracy, which we want to know. To see this, define C^i as the event where classifier/expert i is correct[1]. Then, notice that

$$\begin{aligned} P(f^i(x) = f^j(x)) &= P(C^i \cap C^j) + P(\overline{C}^i \cap \overline{C}^j) \\ &= P(C^i \cap C^j) + P\left(\overline{C^i \cup C^j}\right) \\ &= P(C^i \cap C^j) + 1 - P(C^i) - P(C^j) + P(C^i \cap C^j) \\ &= 1 - \alpha^i - \alpha^j + 2\alpha^{\{i,j\}}, \end{aligned}$$

where $\alpha^{\{i,j\}}$ is the "joint accuracy" of classifier/expert i and j.

This relationship between agreement and accuracy for pairs of classifiers/experts defines $\binom{E+1}{2}$ equations in $E + 1 + \binom{E+1}{2}$ unknowns. Since this is an underdetermined system of equations, the authors suggest minimizing an objective function with the agreement equations as constraints. Specifically, they minimize

$$O(\boldsymbol{\alpha}) = \sum_{\substack{i,j=0 \\ i \neq j}}^{E} \left(\alpha^i \alpha^j - \alpha^{\{i,j\}}\right)^2.$$

The effect of this objective is to find accuracies that satisfy the constraints and imply the least amount of dependence possible.

[1] We denote the event where classifier/expert i is incorrect as $\bar{C}^i$

Notes

1. Platanios et al. [3] present their method for the binary case. We will thus show results only on binary problems.

2. Higher-order agreement rates can be related to functions of higher-order accuracies and thus incorporated into the optimization problem as additional constraints. However, previous work has found that that this has little benefit in terms of the quality of accuracy estimation and makes the optimization problem much more difficult to solve. Thus, in our implementation, we consider only pairwise agreement rates.

3. Platanios et al. [3] derive additional constraints to add to the optimization problem using some simple rules of probability. To be specific, since

$$P(C^i \cap C^j) = P(C^i \,|\, C^j)P(C^j) \leq P(C^j) \quad \text{and} \quad P(C^i \cap C^j) = P(C^j \,|\, C^i)P(C^i) \leq P(C^i)$$

   we have that

$$\alpha^{\{i,j\}} \leq \alpha^i \quad \text{and} \quad \alpha^{\{i,j\}} \leq \alpha^j.$$

4. Notice that the derivations above can also be derived in terms of error rates, rather than accuracy. As a consequence, it is not possible to distinguish between a classifier/expert that is 10% accurate and a classifier/expert that is 90% accurate. To rectify this identifiability problem, Platanios et al. [3] introduce the constraints $\alpha^i \in (0.5, 1]$, which encodes the assumption that the classifiers are all better than random.

3.2.2 Agreement between expert consensus and classifier (AGR)

Lehner [5] derives an estimator that is similar to that discussed above. However, Lehner isolates the classifier from the experts and also assumes that all experts have the same symmetric confusion matrix. That is, given the true class, each expert is right with probability α^e and chooses one of the remaining $L-1$ classes with probability $(1-\alpha^e)/(L-1)$. The classifier is assumed to have a symmetric confusion matrix with parameter α^0.

Lehner uses agreement between the experts to estimate α^e, the expert labels to estimate the base rate of the classes, and then Bayes' rule to estimate the probability that an instance belongs to each class. Finally, assuming these probabilities represent the probability that each label is correct, let f^{hp} be the class label with the highest probability and α^{hp} be the probability assigned to that label. Let C^{hp} be the event that the class label with the highest probability is correct. Then, using the law of total probability, Lehner shows that

$$\begin{aligned} P(f^0(x) = f^{hp}(x)) &= P(f^0(x) = f^{hp}(x) \,|\, C^{hp})P(C^{hp}) + P(f^0(x) = f^{hp}(x) \,|\, \overline{C}^{hp})(1 - P(C^{hp})) \\ &= \alpha^0 \alpha^{hp} + \frac{1-\alpha^0}{l-1}(1-\alpha^{hp}), \end{aligned}$$

assuming that the classifier label is independent of the most probable label. It is now possible to solve for α^0

$$\alpha^0 = \frac{(L-1)P(f^0(x) = f^{hp}(x)) + \alpha^{hp} - 1}{L \cdot \alpha^{hp} - 1}.$$

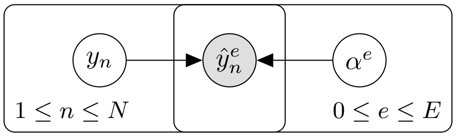


Figure 1. Graphical model of label generation.

Agreement rates are measured by binning the data instances by expert label probability and then determining the proportion of instances in each bin where the most probable label and the classifier label agree. Then, the bin-wise α^0 are computed and combined via a weighted average.

3.3 GRAPHICAL MODELS (MLE AND BEE)

Two groups of authors, Donmez et al. [2] and Platanios et al. [6], consider a graphical model that describes classifier/expert label generation and use this to infer accuracy. We display the graphical model in Figure 1. The nodes y_n, $1 \leq n \leq N$ represent the (unknown) true label for each data instance. The nodes α_e, $0 \leq e \leq E$ represent the (unknown) accuracy of classifier/expert e, while the nodes $\hat{y}_n^e$ represent the label of the e^{th} classifier/expert on the n^{th} data instance. Donmez et al. [2] approximate the maximum likelihood estimate of accuracy using Expectation-Maximization. Platanios et al. [6] set priors for y_n and α^e and use Gibbs sampling for inference.

Notes

1. Platanios et al. [6] present their method for the binary case. We extended the method to apply to multi-class problems, but found the inference procedure to work inconsistently in this case. We will thus show results only on binary problems.
2. Many other authors have proposed related graphical models for the purpose of inferring the true label [9–16] many of which contain accuracy based parameters and could thus be used to estimate accuracy as well. These approaches typically incorporate additional effects such as item difficulty [11] or expert dedication [13]. We consider the simplest versions of these models because previous work has found them to perform well in practice.

3.4 COVARIANCE-BASED ESTIMATOR (COV)

Parisi et al. [17] and Jaffe et al. [4] use the labeler covariance matrix to either rank a group of binary classifiers [17] or estimate their class conditional accuracies [4].

Parisi et al. [17] show that the off-diagonal entries of the classifier covariance matrix are equal to the off-diagonal entries of a rank-one matrix S. That is:

$$\underbrace{\begin{bmatrix} \quad \mathbf{E}\left[(f^i-\mu^i)(f^j-\mu^j)\right] \quad \\ \end{bmatrix}}_{\Sigma} \underbrace{=}_{\forall i\neq j} \underbrace{\begin{bmatrix} \quad S_{ij} \quad \\ \end{bmatrix}}_{S} = \underbrace{\begin{bmatrix} s_i \\ \end{bmatrix} \begin{bmatrix} \quad s_i \quad \end{bmatrix}}_{ss^T}.$$

In addition, they show that

$$s_i = \sqrt{1-b^2}(\psi^i_{11} + \psi^i_{22} - 1), \tag{1}$$

where $b = \pi_1 - \pi_2$, and that the classifier means μ^i can be written as

$$\mu^i = (\psi^i_{11} - \psi^i_{22}) + b(\psi^i_{11} + \psi^i_{22} - 1). \tag{2}$$

Solving the system of linear equations from (1) and (2) in terms of μ^i, s_i, and b, we reach

$$\psi^i_{11} = \frac{1}{2}\left(1 + \mu^i + s_i\sqrt{\frac{1-b}{1+b}}\right)$$

and

$$\psi^i_{22} = \frac{1}{2}\left(1 - \mu^i + s_i\sqrt{\frac{1+b}{1-b}}\right).$$

The means μ^i and also b can both be easily estimated from the classifier labels. Computing the diagonal of S is more difficult and Parisi et al. [17] discuss several algorithms to do so. Once this is done, we can recover s from the singular value decomposition of S.

Notes

1. Jaffe et al. [4] extend the work of Parisi et al. [17] and show how to use the "covariance tensor"

   $$\mathbf{E}\left[(f^i-\mu^i)(f^j-\mu^j)(f^k-\mu^k)\right]$$

   to consistently estimate b. We find that for small sample sizes, estimating b simply as the proportion of classifier and expert labels often performs better.

2. This approach extends easily to multi-class problems by considering a sequence of one vs. all estimation problems. The estimate to sensitivity for each binary problem becomes the estimated diagonal entry of the confusion matrix.

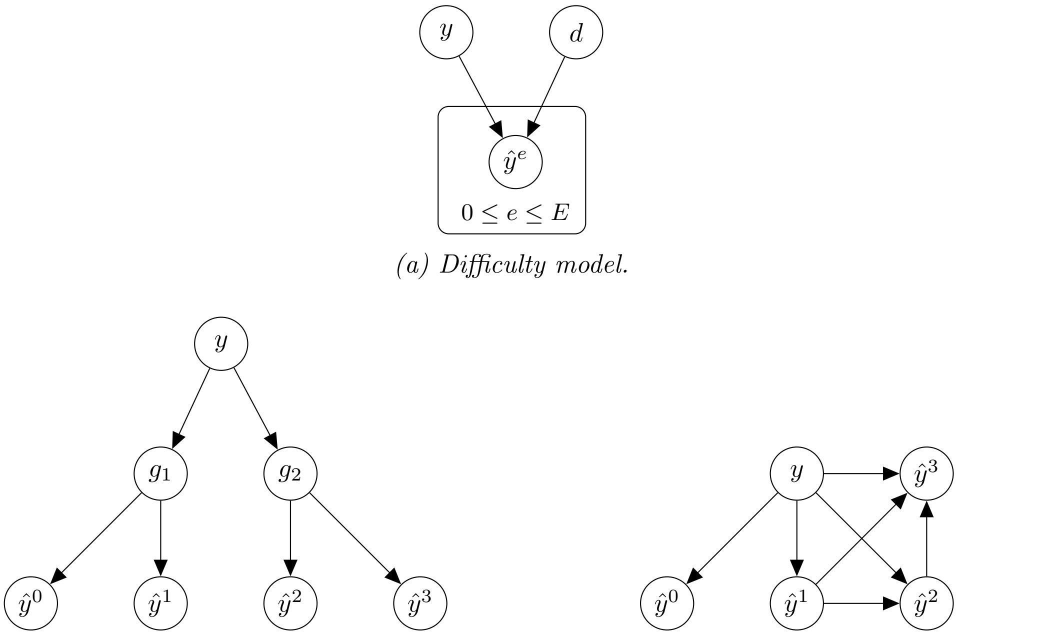
4. SIMULATIONS

In this section, we consider three models of label generation (difficulty, groups of experts, and copy) and then examine the results of applying the estimators described in Section 3 to datasets simulated from the models. Each model simulates conditional dependence between experts in a different way. Two of the models, groups of experts and difficulty, appear in the literature (see [18] and [19]). The copy model is our own construction. We use the models to address the following questions:

1. Is there a relationship between conditional dependence and estimator error?
2. If expert and classifier accuracies are different from one another, what effect does this have on estimator error?
3. Can we draw broad inferences about the performance of the estimators relative to one another?

4.1 MODELS CONSIDERED

Difficulty model The first model of expert dependence, depicted in Figure 2a, assumes that in a dataset, some instances are inherently more difficult than others. True labels are drawn from a Dirichlet class label prior distribution $\pi = \mathsf{Dir}(1, \ldots, 1)$. We consider the case where instances are



(a) Difficulty model.

(b) Groups of experts model with two groups.

(c) Notional copy model for three experts.

Figure 2. Models of label generation.

either “easy” or “hard” (encoded in the model by the d node). Experts choose their label from a different confusion matrix depending on whether the instance is easy or hard. Clearly, the experts are no longer conditionally independent, given only the true label.

Groups of experts model The second model of dependence assumes that the experts are divided into groups and choose a label for each data instance after viewing the true label through their group's “intermediate node.” The model is depicted in Figure 2b. Notice that experts in the same group are conditionally dependent, given the true label, but that experts in different groups are conditionally independent. It should be noted that this notion of dependence is global in nature. If two experts fall into the same group, they are dependent throughout the simulation.

In this simulation, the true label y is sampled from a Dirichlet class label prior distribution $\pi = \mathsf{Dir}(1, \ldots, 1)$. Next, a noisy version of the label g_k is generated for each group k, according to the group confusion matrix ϕ_k. The experts view g_k and then generate a label according to their confusion matrix ψ^e.

Copy model The third model induces dependence between experts via a random instance-dependent partition of experts. The assignment of experts to groups is similar to the Chinese Restaurant Process (CRP) [20].

Formally, for a single instance, the true label y is sampled from a Dirichlet class label prior distribution $\pi = \mathsf{Dir}(1, \ldots, 1)$. The classifier selects a label using its own confusion matrix. The first expert's label is generated in an identical fashion using the expert confusion matrix. Given that e expert labels have been produced, the $e+1$-st is generated by the following process: with probability ρ, expert $e+1$ copies the label of a previous expert (chosen uniformly at random) and with probability $1-\rho$, expert $e+1$ will generate a label independent of previous experts using the expert confusion matrix. Note that this model differs from the CRP in that the probability of generating a new label is fixed at $1-\rho$, whereas in the CRP this probability is inversely proportional to the number of experts who have already chosen a label. Also note that for $\rho = 0$, the experts are conditionally independent and when $\rho = 1$, all experts provide the exact same set of labels. A notional depiction of the model is provided in Figure 2c.

The motivation for this model is twofold. In early simulations, it was quickly observed that as expert accuracy decreased, error in estimating the classifier's accuracy increased. Additionally, many of the estimators either assumed conditional independence of the experts, or proved performance guarantees with a conditional independence assumption. The copy model allows us to easily decouple accuracy and dependence (as compared to the difficulty and groups of experts models). In contrast to the other models, there is no notion of global dependence; experts are dependant only on a per-instance basis.

4.2 MODEL PARAMETERIZATION

We performed simulations across the three expert labeling models using parameter grids. In all simulations, we fixed the number of experts at $E = 5$, the number of classes at $L = 5$, the number of samples at $N = 100$, and considered 100 Monte Carlo replicates. We now discuss the values considered for each model's parameters.

Difficulty model In order to ensure that the "hard" instances in the difficulty model are always more challenging to the simulated experts and classifiers in the difficulty error model, we set accuracy on the hard instances as a percentage penalty of the accuracy on the easy instances. For example if the expert accuracy on easy instances is 0.8 and the hard penalty is 0.4, then expert accuracy on hard instances is given by $0.8 \cdot (1 - 0.4)$. We considered accuracies of between 0.4 and 0.9, in increments of 0.1 on the easy instances for both the experts and classifier, and hard penalties of between 0.1 and 0.5 in increments of 0.1. We set the probability that an instance is easy to 0.5. As a result, the accuracy of a classifier or expert is $\frac{x}{2} + \frac{x(1-y)}{2}$, where x is the classifier/expert accuracy on easy instances and y is the hard penalty.

Groups of experts model For the groups of experts model, we let the number of groups vary from one to six, fixed the "noise rate" of the intermediate nodes (g_k) at 0.8, and considered "noisy accuracies" of between 0.4 and 0.9 in increments of 0.1 for the experts and classifier. By noisy accuracies, we mean the probability with which the classifier or expert labels match the intermediate node.

Copy model For the copy model, we used accuracies of 0.4 and 0.9 in increments of 0.1 for the experts and classifier and used expert correlations ρ between 0 and 0.8 in increments of 0.1.

4.3 RESULTS

In the rest of this section, we will address the three questions we outlined earlier; whether conditional dependence affects estimator error, how differences in classifier and expert accuracy affect estimator error, and whether we can draw any broad conclusions about the performance of the estimators relative to one another.

Expert Dependence To address the issue of how conditional dependence between experts affects estimator error, we will use the groups of experts model and copy model. In both of these models, we can vary the overall "amount" of dependence without affecting expert or classifier accuracy. We cannot easily do this in the difficulty model, so we do not present results for that here. In the groups of experts model, note that increasing the number of groups causes more pairs of experts to be conditionally independent. As a result, we expect that as the number of groups increases, the error of the estimators will decrease. In the copy model, increasing the copy chance introduces additional dependence between the experts and so we expect that the error of the estimators will increase.

In Figure 3, we show the average absolute errors and one standard deviation error bars for the estimators applied to the groups of experts model. In each plot, the horizontal axis contains

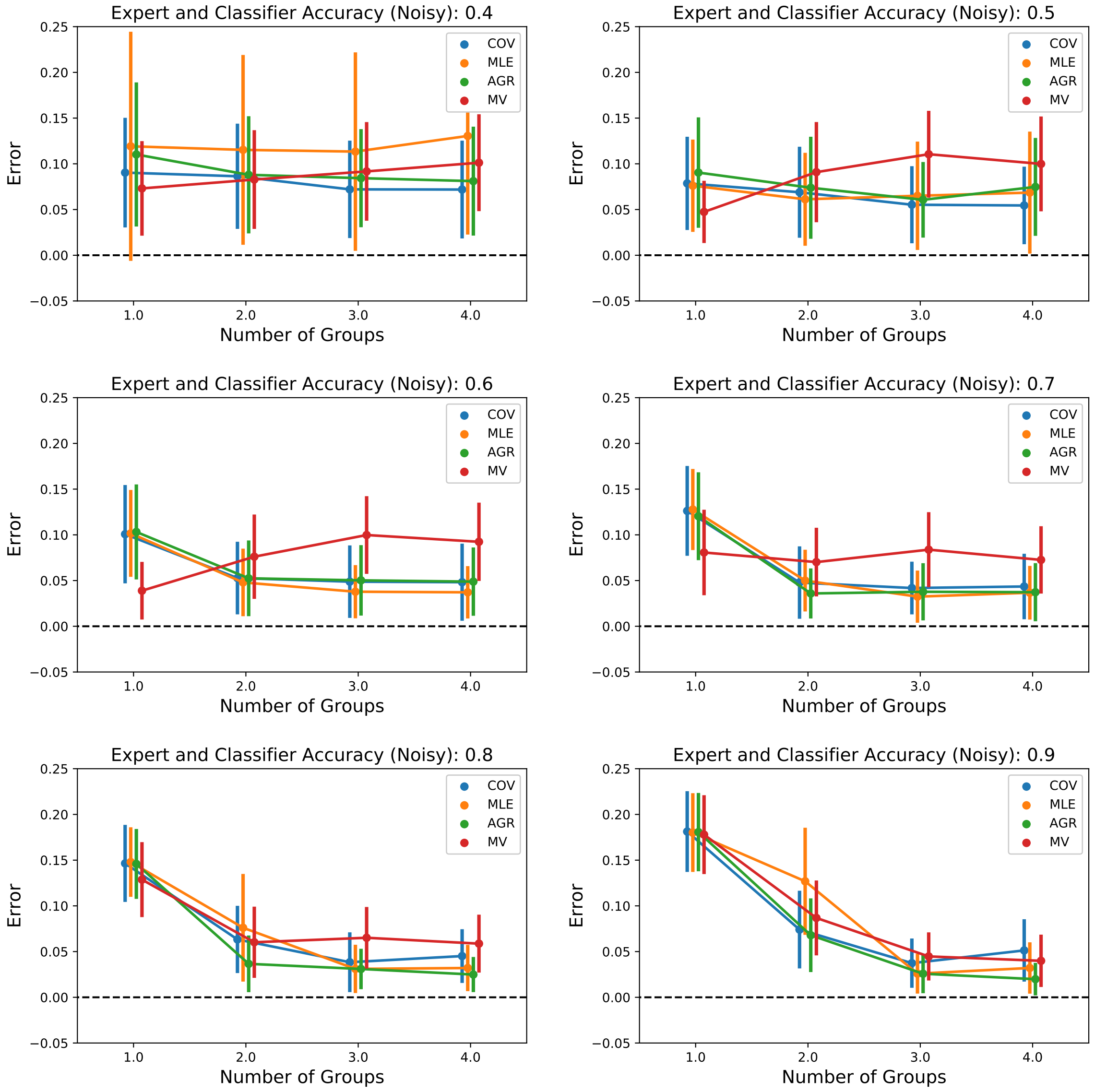


Figure 3. Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots. The actual accuracies are 0.8 *times the noisy accuracies.*

the number of groups. Recall that as the number of groups increases, overall dependence between experts decreases. The difference between the plots in the figure is the noisy accuracy of the experts and classifier. We note that when expert and classifier noisy accuracy is either 0.7, 0.8, or 0.9, there is a obvious decrease in error as the number of groups increases for COV, MLE, and AGR, especially between one and two groups. When expert and classifier accuracy is low (especially at 0.4), there is not as clear of a decrease in estimator error. It appears then that the effect of dependence on estimator error is more pronounced as the experts and classifier become more accurate.

Figure 4 shows average absolute errors and one standard deviation error bars for the copy model of dependence. In these plots, the horizontal axis shows $1 - \rho$. Recall that as ρ increases, dependence increases, or as $1 - \rho$ increases, dependence decreases. Across expert and classifier accuracies, the overall trend is clear, as dependence between experts decreases, the error of the estimators decreases. The trend is much more consistent across accuracies and estimators than for the groups of experts model.

Expert and Classifier Accuracy We also want to examine the relationship between expert and classifier accuracy and estimator error. To do this, we will plot estimator error against expert accuracy, for fixed values of expert dependence and classifier accuracy.

In Figure 5, we show results for the difficulty model. In all the plots, the hard instance penalty is 0.2. Our primary observation is that (except when classifier accuracy is 0.4) MLE, AGR, and COV generally outperform MV when experts have low accuracy. It also appears that the difference between MV and the others becomes larger as the classifier accuracy increases. For other hard instance penalties, we observed a very similar pattern and as a result do not display those plots here.

In Figure 6, we show results for the copy model. In this model, we again see that COV, MLE, and AGR outperform MV when experts have low accuracy. As the classifier accuracy improves, the performance of all estimators get worse and so the difference between MV and the others does not increase as obviously as in the difficulty model. In these plots, we are showing results for $\rho = 0.2$. As ρ increases, the differences between the estimators becomes smaller. For $\rho = 0.0$ or $\rho = 0.1$, the differences between MV and the others is sharper.

Finally, in Figures 7 and 8, we show results for the groups of experts model with one and three groups, respectively. When there is only one group (see Figure 7), all of the estimators perform poorly. In fact, we see no improvement as the experts become more accurate. The u-shaped curve for MV appears odd at first, but is explained by looking at the signed errors. MV tends to underestimate classifier accuracy when experts have low accuracy and then overestimate classifier accuracy when experts have high accuracy. In the between the two extremes, it crosses the true accuracy of the classifier, which is why MV has small absolute errors for some expert accuracies. These results seem to indicate that the groups of experts model with one group is particularly adversarial to the estimators. When there are three groups (see Figure 8), we again see that MLE, COV, and AGR have a clear advantage over MV when experts have low accuracy. For two groups, results are similar to those for one. For four, five, or six groups, results are similar to those for three groups.

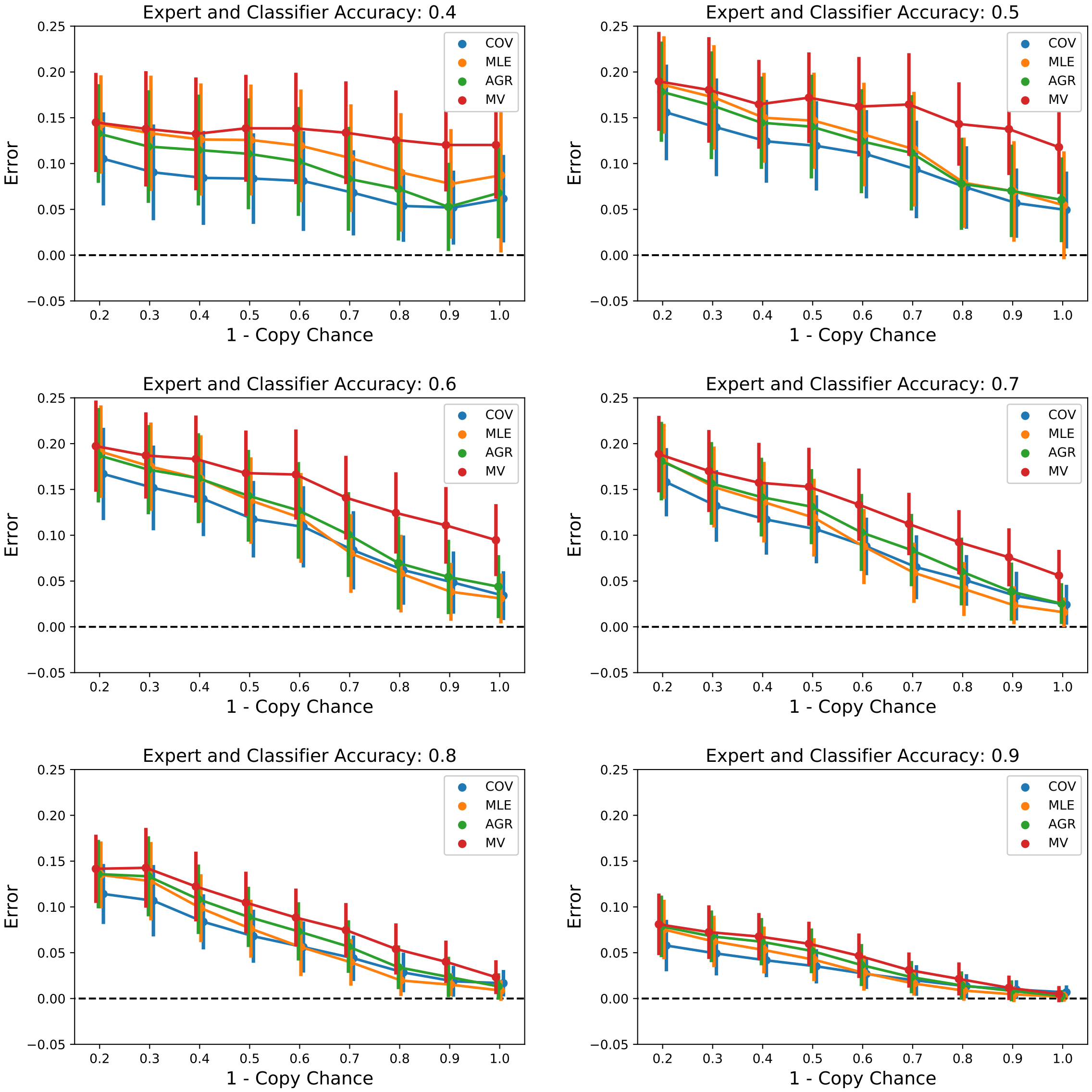
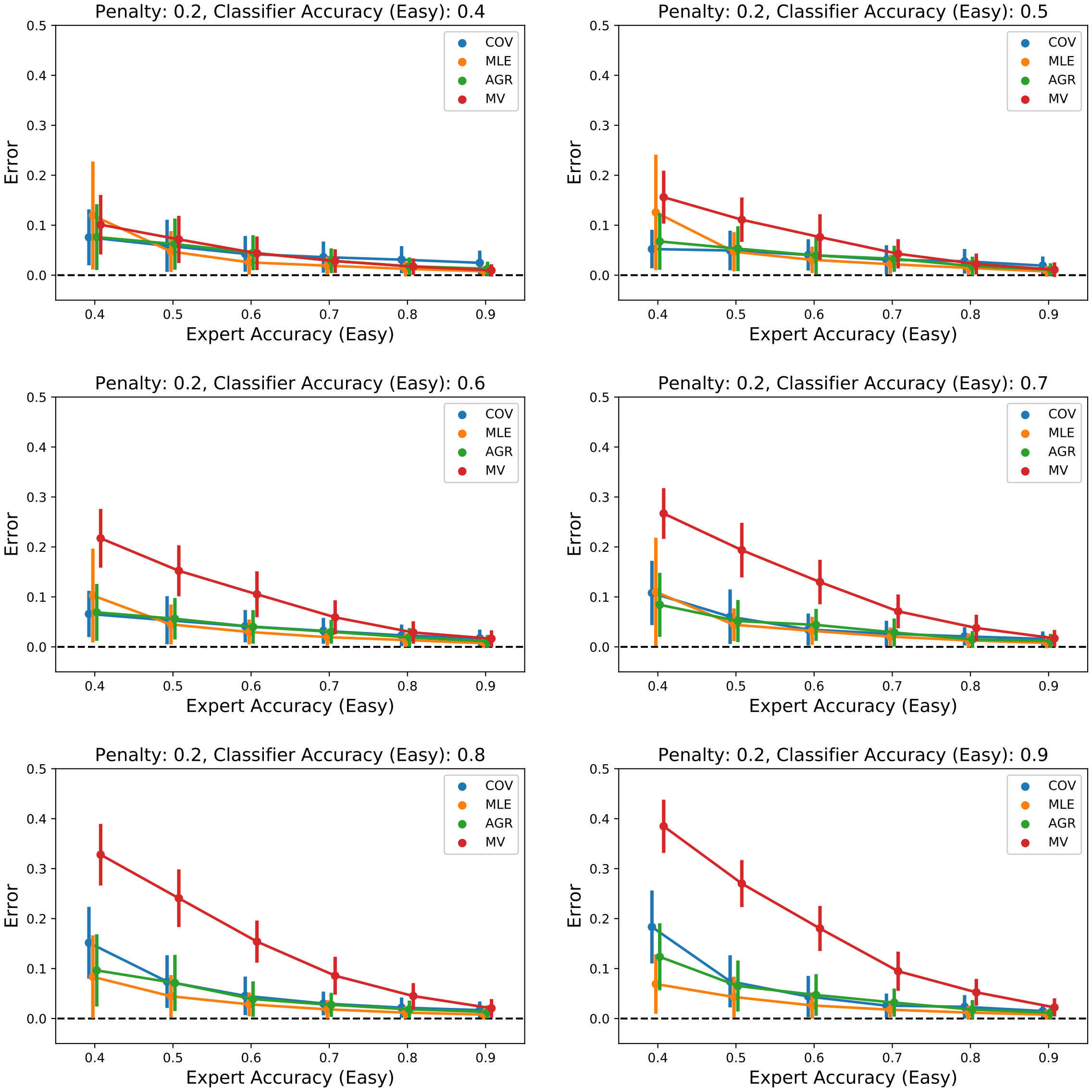


Figure 4. Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the copy model of dependence for each estimator. The accuracies of the experts and classifier are given in the title of the individual plots.



Penalty: 0.2, Classifier Accuracy (Easy): 0.4

Penalty: 0.2, Classifier Accuracy (Easy): 0.5

Penalty: 0.2, Classifier Accuracy (Easy): 0.6

Penalty: 0.2, Classifier Accuracy (Easy): 0.7

Penalty: 0.2, Classifier Accuracy (Easy): 0.8

Penalty: 0.2, Classifier Accuracy (Easy): 0.9

Figure 5. Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the difficulty model of dependence for each estimator. The accuracy penalty on hard instances and the accuracy of the classifier on easy instances are given in the title of the individual plots. The overall accuracy of the experts or classifier is $\frac{x}{2} + \frac{x(1-y)}{2}$*, where* x *is the accuracy on easy instances and* y *is the hard instance penalty.*

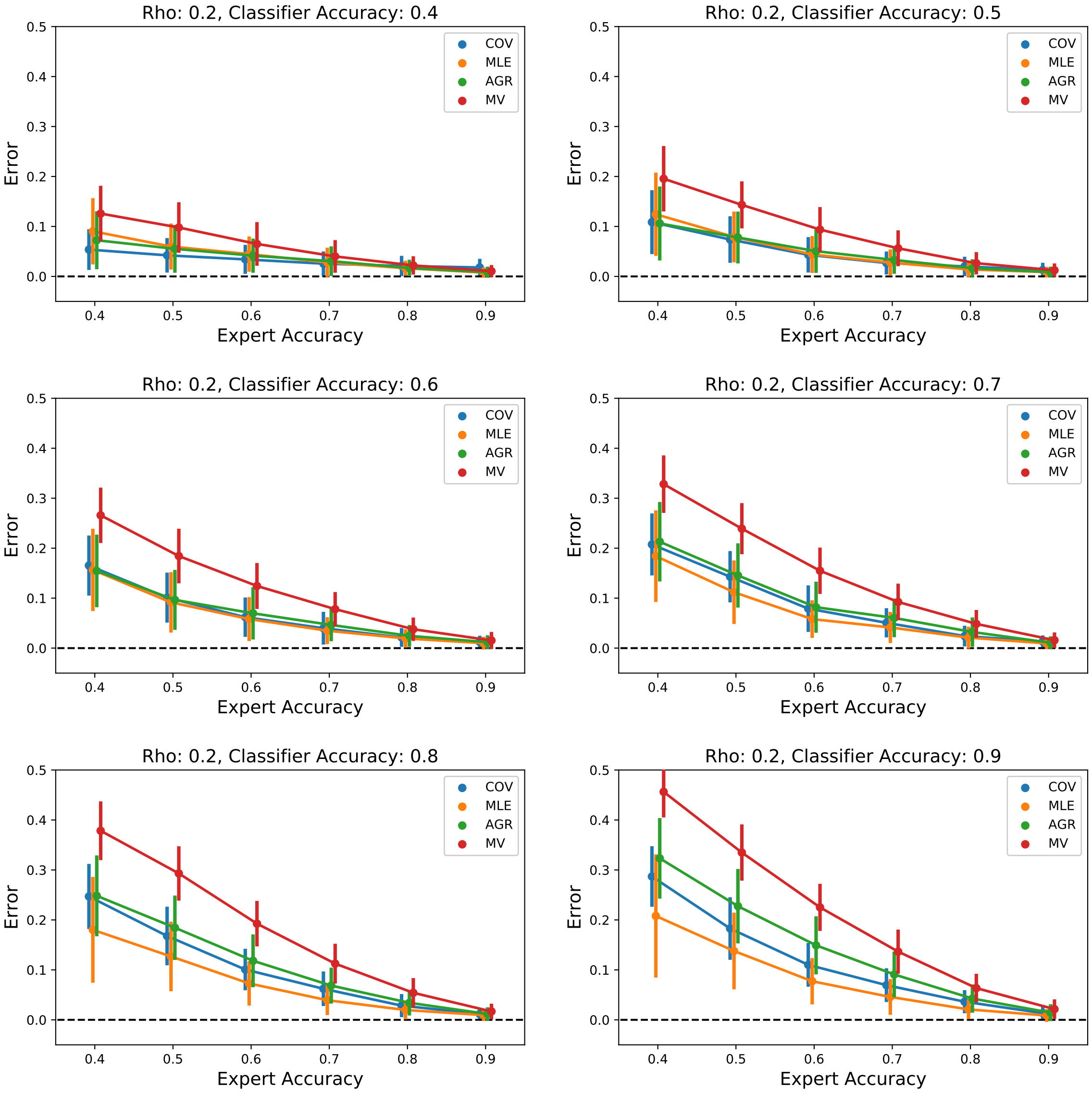


Figure 6. Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the copy model of dependence for each estimator. The value of ρ and the accuracy of the classifier are given in the title of the individual plots.

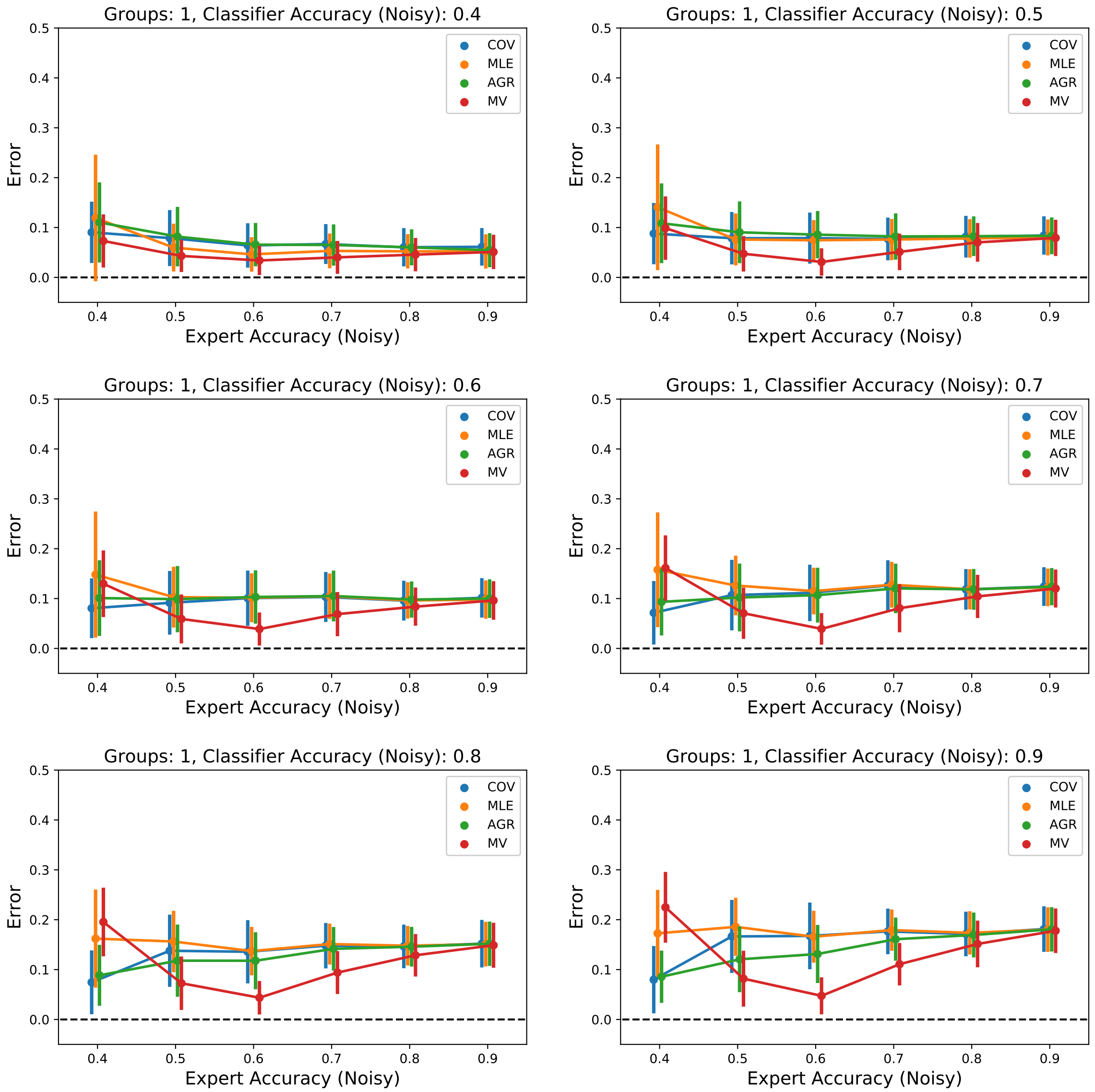
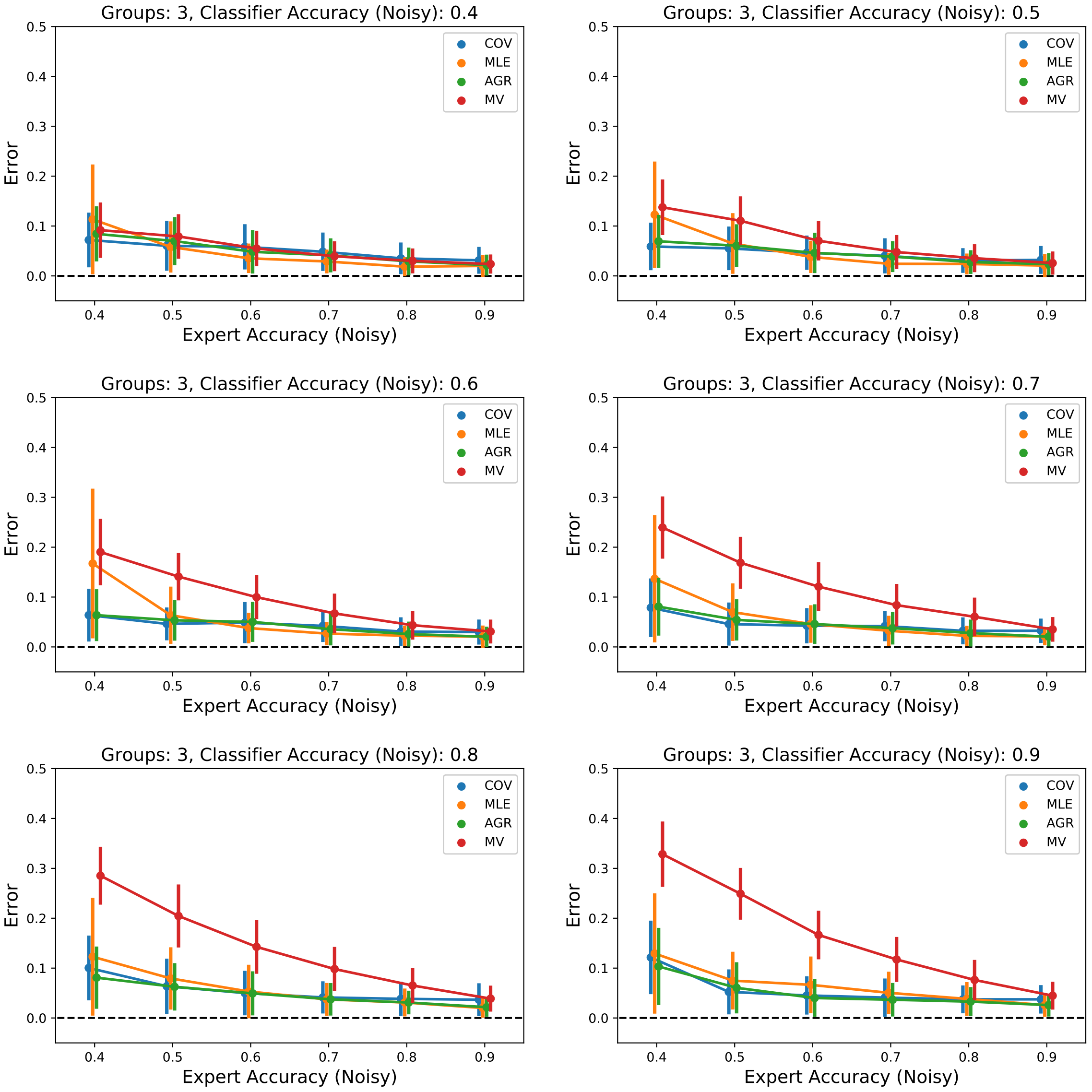


Figure 7. Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence with one group for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots. The actual accuracies are 0.8 *times the noisy accuracies.*



Groups: 3, Classifier Accuracy (Noisy): 0.4

Groups: 3, Classifier Accuracy (Noisy): 0.5

Groups: 3, Classifier Accuracy (Noisy): 0.6

Groups: 3, Classifier Accuracy (Noisy): 0.7

Groups: 3, Classifier Accuracy (Noisy): 0.8

Groups: 3, Classifier Accuracy (Noisy): 0.9

Figure 8. *Average absolute errors and one standard deviation error bars for 100 Monte Carlo samples per parameter choice in the groups of experts model of dependence with three groups for each estimator. The noisy accuracies of the experts and classifier are given in the title of the individual plots.*

Relative Performance of the Estimators From the results displayed in this section, it seems clear that MLE, COV, and AGR have some advantage over MV, particularly when the experts have low accuracy and the classifier has high accuracy. Between MLE, COV, and AGR, there does not seem to be a clear best performer. They tend to perform very similarly overall, and none is consistently better than the others, across the simulations.

It is disappointing but perhaps not surprising that none of the MLE, COV, or AGR appears to handle dependence between experts better than another. We argue that this is not surprising, as despite being very different in their construction, each assumes conditional independence between experts.

4.4 SUMMARY

We used three models to generate noisy expert labels that are not necessarily conditionally independent to isolate several potential issues with using noisy labels to estimate accuracy. These are the copy model, which on a per-instance basis generated clusters of experts who all chose the same label and between clusters were conditionally independent; the groups of experts model, which placed experts into groups, wherein experts received the same noisy observation of the true label; and the difficulty model, which introduced heterogeneity into the accuracy of experts. From these simulations, we reach the following conclusions:

1. Increasing the degree of dependence between experts increases error in the copy model and either increased or left unchanged the error in the groups of experts model.
2. As the accuracy of the classifier increases relative to the accuracy of the experts, the performance of all estimators suffer. However, COV, AGR, and MLE tended to be more robust to these effects than MV.
3. MV tended to perform poorly when compared with COV, AGR, and MLE. In general, there is no clear favorite between COV, AGR, and MLE; for some simulations, one outperforms the others and vice versa.

This page intentionally left blank.

5. EXPERIMENTS

We discuss four experiments (Sections 5.1–5.4) conducted over the course of this study. In each experiment, we gathered a labeled dataset and divided it into a training and testing set. We then trained a classifier on the training set and applied it to the test set. Our goal is to estimate the performance of the classifier on the test set (without using the true labels). To do this, we sample a small number of data instances from the test set (100 or 200) and ask several human experts to label each instance. Using the labels of the classifier and the experts on this sample, we apply the estimators from Section 3 to estimate accuracy.

In each of the sections below, we will describe the specific experiment and show the performance of the estimators in terms of signed error over bootstrap resamples of the data and in terms of absolute error as the number of experts is increased. We will also examine the degree of dependence between the classifier and experts. We defer discussion regarding *detecting* dependence between experts until Section 6.

5.1 JOB CATEGORY

Dataset In this experiment, we used a dataset taken from a medium-sized enterprise network of about 2000 users. The features were taken from two separate three-month periods, January–March 2016 and April–May 2016; these two periods define our training and test dataset. The classification task is to determine to which job category (Staff, Administrator, or Management) users belong. The features are as follows:

- Years employed, whether user has organization-owned mobile device
- Number of weekdays with no logins during normal work hours, number of weekdays with a login
- Average time per week on VPN (morning, work hours, night), number of days with a VPN connection from out of state
- Number of emails sent, number of emails received, number of users to whom email was sent, number of users from whom email was received

The classification task is non-trivial as users across classes will have similar behavior for some features. For example, both the Administrator and Management classes may send many emails, while both Staff and Management may regularly connect via VPN or have an organization-owned mobile device.

Classifier and experts For our classifier, we used a random forest. We collected labels on 100 samples from each of nine experts. On the left of Figure 9, we show the accuracies of each of the nine experts on the sample and the accuracy of the classifier on the entire test dataset. The classifier's accuracy is about 88%, while the experts' accuracies ranged from 45–83%. On the right

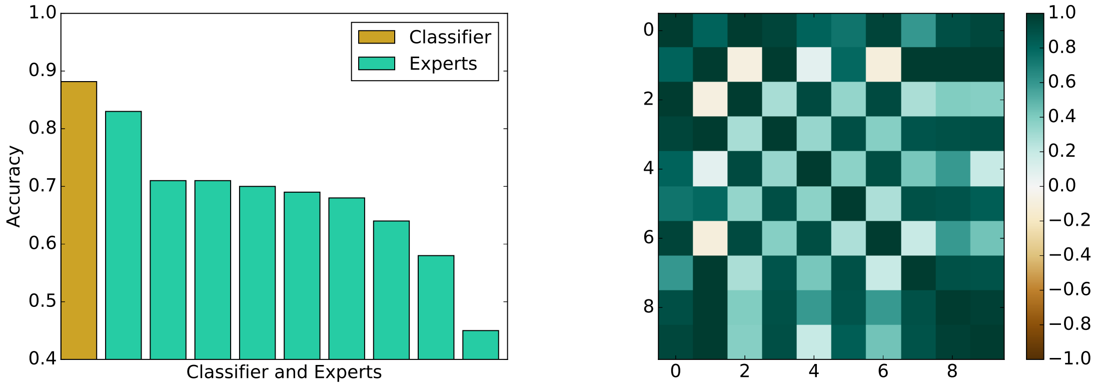


Figure 9. Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.

of the figure, we show a class-averaged version of Yule's Q statistic, which measures the degree of conditional dependence between experts. Values close to 1 indicate a large positive conditional dependence, while values close to -1 indicate a large negative dependence. On the other hand, when the value is 0, the experts are conditionally independent. It is clear that many experts exhibit significant positive dependence. The checkerboard pattern indicates there may be structure in the dependence relationships; we defer this discussion until Section 6.

Estimator errors Having described the experiment and the performance of the classifier and experts, we now apply the estimation methods to estimate the accuracy of the classifier. To get a sense for the variance of the estimators, we resampled the dataset 10,000 times, applied the estimator to each resample, and computed the errors. We show boxplots of the errors (and the individual errors) for each estimator in Figure 10. We also include the so-called empirical estimator EMP, which is simply the proportion of labels that are correct (using the known ground truth). For a given sample of data, estimating accuracy on the test set as the proportion correct in the sample is the best we can do.

It is clear that, overall, the estimators tend to underestimate accuracy. Specifically, the median error is negative for MLE, COV, AGR, and MV. MLE provides the best median performance but seems to have higher variance than the other estimators. While the estimates appear biased, they do tend to be reasonably good approximations. No estimator ever had an absolute error greater than 0.30 and the median error of each estimator is less than about 0.10.

We considered testing for differences in the mean *absolute* errors of the estimators. However, due to the large number of resamples that we performed, even very small differences in the mean absolute errors can be statistically significant. To avoid confusing the issue, we simply display the mean absolute errors of the estimators in Table 1. MLE appears to perform the best with a mean absolute error of 0.065, while MV performs the worst with a mean absolute error of 0.100.

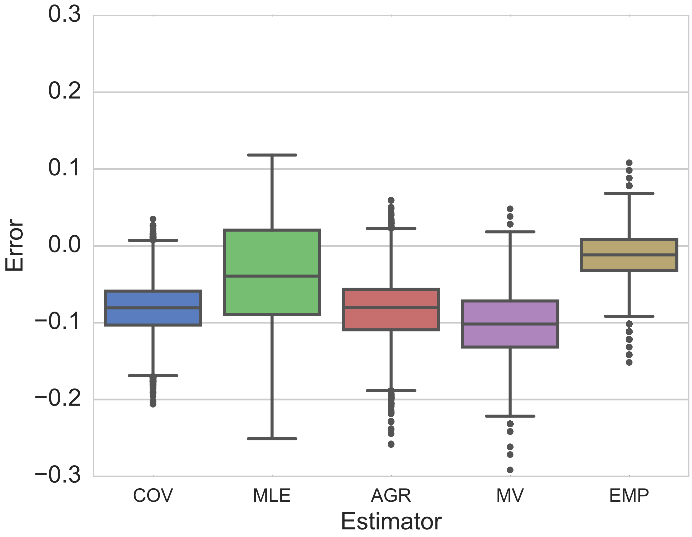


Figure 10. Signed errors for 10,000 resamples of the Job Category dataset for each estimator.

Increasing the number of experts We are also interested in whether using more experts in the estimators decreases *absolute* error. To answer this question, we apply each estimator to every possible subset of experts of size $2, 3, \ldots, E-2$. In Figure 11, we show boxplots of the results for MLE and AGR. For MLE, the median error declines as the number of experts increases. Regardless of the number of experts, absolute error is nearly always less than 0.10. For AGR, the performance as the number of experts increases is quite inconsistent. It appears that the performance is different for odd and even numbers of experts. In particular, it is better on this dataset for an even number of experts, though the performance gets worse as the number is increased. For odd numbers of experts, the performance remains about the same. We hypothesize that this behavior is due to the fact that it chooses a single most probable label to use in the agreement with the classifier calculation. We observed similar behavior with MV, which also chooses a single consensus label to compare against the classifier's.

TABLE 1

Mean absolute errors of the estimators across 10,000 resamples of the Job Category dataset.

Estimator	Mean Absolute Error
COV	0.081
MLE	0.065
AGR	0.083
MV	0.100

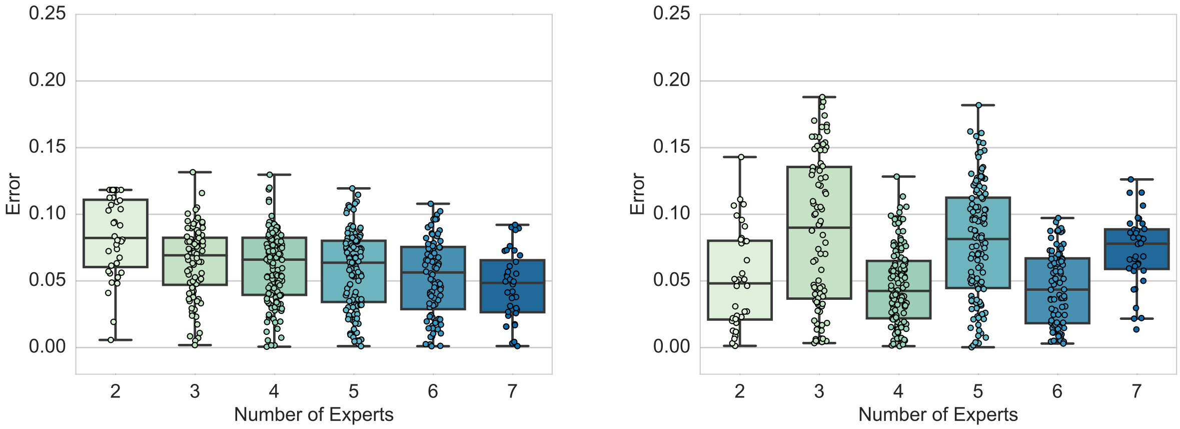


Figure 11. Absolute errors of MLE *(left) and* AGR *(right) for all possible subsets of between two and seven experts, on the Job Category dataset.*

5.2 CYBER SOCIAL MEDIA CONVERSATIONS I

Dataset In this experiment, we gathered social media conversations from three StackExchange forums. StackExchange is a website divided into topic-based forums where users can ask questions. We collected conversations from the Crypto, Security, and ReverseEngineering forums (158, 129, and 316 conversations, respectively). We used half the dataset for training and half for testing.

Similar topics are discussed in each forum, and it is sometimes difficult to tell the true class. For example, the following two conversations highlight a case that is not immediately clear:

1. I was thinking recently about password security. My goal is to have mostly random passwords, that are different for each site. But you also should be able to remember them (or re-generate them) without the help of any notes or the like...
2. I am currently reading about PBKDF2, and understand that the salt is used only once, while the password is used multiple times in the computation of the final key (see this question). How would the integrity of PBKDF2 change if the roles of password and salt are changed?...

While both quotes discuss passwords, the first concerns the writer's opinion on password policies and is from the Security forum, while the second is a highly technical conversation about a key derivation algorithm and is from the Crypto forum.

Classifier and experts For our classifier, we used a support vector machine with a linear kernel. We trained the classifier on term frequency-inverse document frequency (TF-IDF) features. We collected labels on 100 samples from each of eight experts. The experts, of course, examined the raw text, rather than the word count features. On the left of Figure 12, we show the accuracy of each of the experts on the sample and the accuracy of the classifier on the entire test dataset. The classifier's accuracy is about 87%, while the experts' accuracies ranged from 73–83%. On the right of the figure, we show the dependence between experts. Again, we observe moderate to strong positive dependence between experts. Compared to the previous experiment, it is much less obvious whether there is any structure.

Estimator error We again resampled our dataset 10,000 times and computed the error in estimating accuracy for each resampling. In Figure 13, we show a boxplot of errors for each estimator and in Table 2 we show the mean absolute errors. In contrast to the previous experiment, here we see that all estimators have median error that is less than 0.05. In addition, no estimator ever makes an error larger than about 0.20. The mean absolute errors of the estimators are very similar, between 0.029 and 0.038.

Increasing the number of experts We apply each estimator to every possible subset of experts of size $2, 3, \ldots, E-2$. In Figure 14, we show boxplots of the results for MLE and AGR. For MLE, there is a slight decrease in median error as the number of experts increases, and perhaps a small decrease in variance as well. For AGR, we again see that there appears to be a difference in the performance of the estimator for even and odd numbers of experts. For this experiment, odd numbers of experts produce smaller errors than even numbers of experts. Again, the results for MV on this dataset show a similar see–saw type pattern. For both even and odd numbers of experts, increasing the number appears to slightly decrease the median error.

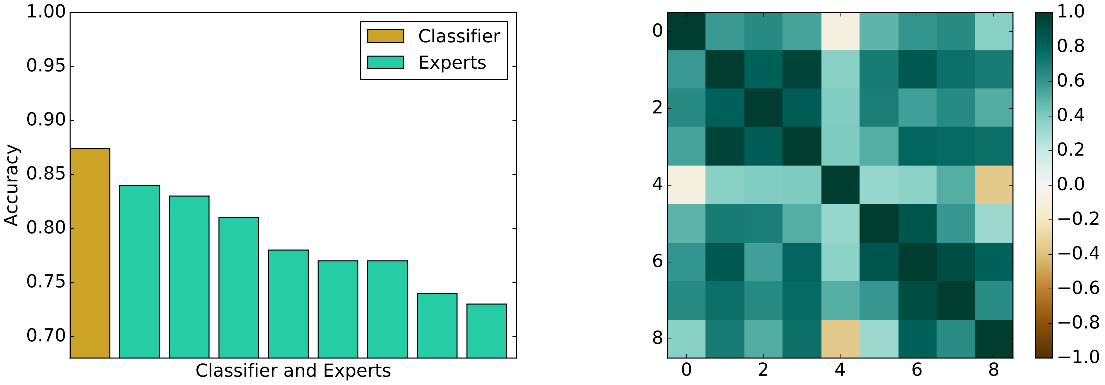


Figure 12. Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.

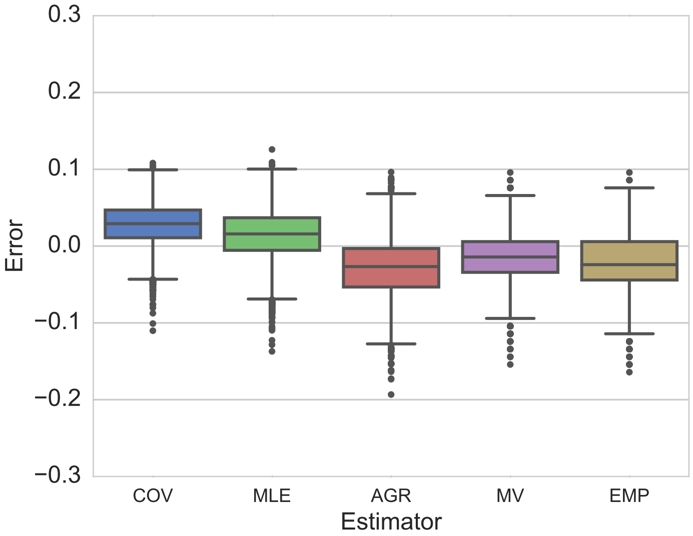


Figure 13. Signed errors for 10,000 resamples of the Social Media I dataset for each estimator.

TABLE 2

Mean absolute errors of the estimators across 10,000 resamples of the Social Media I dataset.

Estimator	Mean Absolute Error
COV	0.033
MLE	0.029
AGR	0.038
MV	0.030

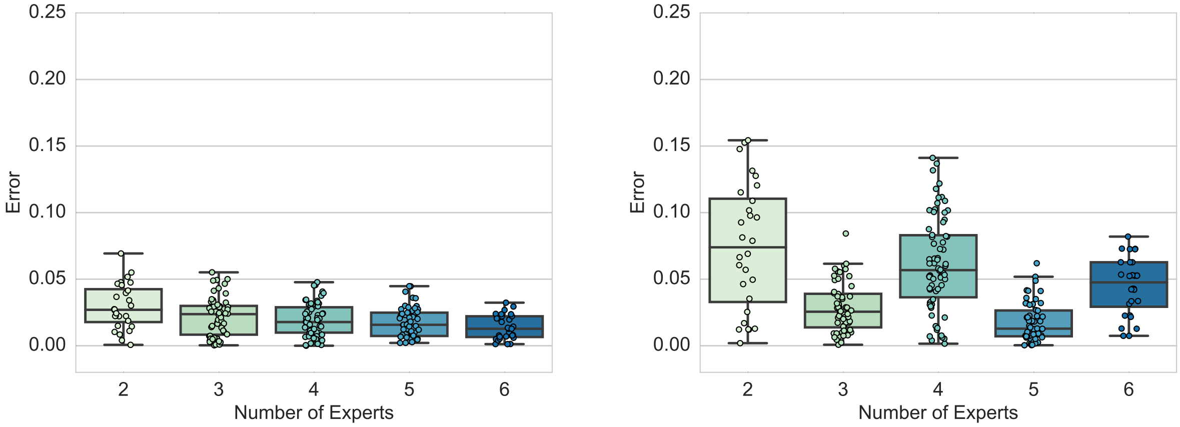


Figure 14. Absolute errors of MLE *(left) and* AGR *(right) for all possible subsets of between two and six experts, on the Social Media I dataset.*

5.3 CYBER SOCIAL MEDIA CONVERSATIONS II

Dataset Our third experiment is very similar to the previous, except that we used conversations from the social media website Reddit, rather than StackExchange. Reddit is also a website that consists of topic-based forums (subreddits), but the conversations tend to be less formal and less focused than those from StackExchange. We collected conversations from the asknetsec, ReverseEngineering, crypto, and darknetplan subreddits (751, 467, 785, and 784 conversations, respectively). We used half the dataset for training and half for testing.

Similar to the example above, the classification is non-trivial for this problem because similar topics may be discussed in different forums. Below are fragments of two conversations in which a user poses a question:

1. How did the bash bug go unnoticed for so long?
   You could ask the same question of Heartbleed. It's an odd bug in an obscure feature of a binary that a lot of researchers don't bother looking at...

2. How to convert huge chunks of data into base N?
   People computing huge values of pi often perform base conversions from hex to decimal for billions or trillions of digits...

While both quotes contain security related questions, the first is a general information question intended to prompt discussion and is from the asknetsec subreddit, while the second is a much more specific question related to cryptography and is from the crypto subreddit.

Classifier and experts For our classifier, we used logistic regression, trained on (TF-IDF) features. We collected labels on 100 samples from each of 8 experts. Again, the experts were shown the raw text, rather than the word count features. On the left of Figure 15, we show the accuracies of each of the experts on the sample and the accuracy of the classifier on the entire test dataset. The classifier's accuracy is about 83%, while the experts' accuracies ranged from 65–81%. On the right of the figure, we show the degree of conditional dependence between classifier/experts. As in the previous experiments, we observe strong positive dependence between classifier/experts. There does not appear to be obvious structure.

Estimator errors We repeat our resampling approach and show the boxplot of errors for each estimator in Figure 16. The results for this dataset are more similar to the first than the second experiment. For this experiment, the estimators are all biased below the true value. Each median error is between -0.05 and -0.10. This bias is perhaps partly explained by the sample, as even EMP is somewhat biased below.

We show the mean absolute errors of the estimators in Table 3. COV performs the best with a mean absolute error of 0.064 while MV has the worst, at 0.108.

Increasing the number of experts We apply each estimator to every possible subset of experts of size $2, 3, \ldots, E - 2$. In Figure 17, we show boxplots of the results for MLE and AGR. For MLE, we see perhaps a small increase in error as the number of experts increases and a small decrease in variance. We see a similar pattern for AGR, that is, roughly similar error as the number of experts increase and slightly smaller variance. Note that the see–saw pattern observed in the two previous datasets does not appear here.

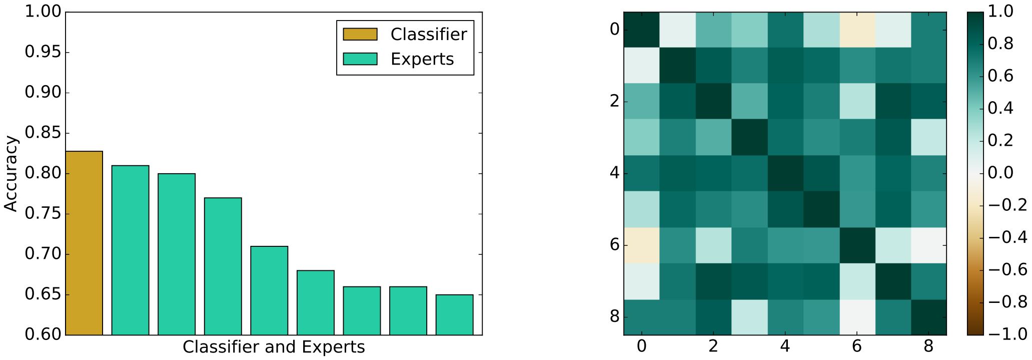


Figure 15. Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.

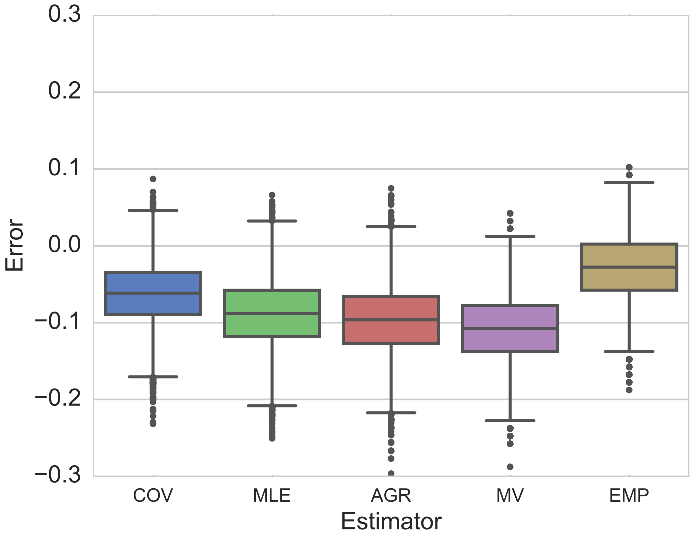


Figure 16. Signed errors for 10,000 resamples of the Social Media II dataset, for each estimator.

TABLE 3

Mean absolute errors of the estimators across 10,000 resamples of the Social Media II dataset.

Estimator	Mean Absolute Error
COV	0.064
MLE	0.089
AGR	0.097
MV	0.108

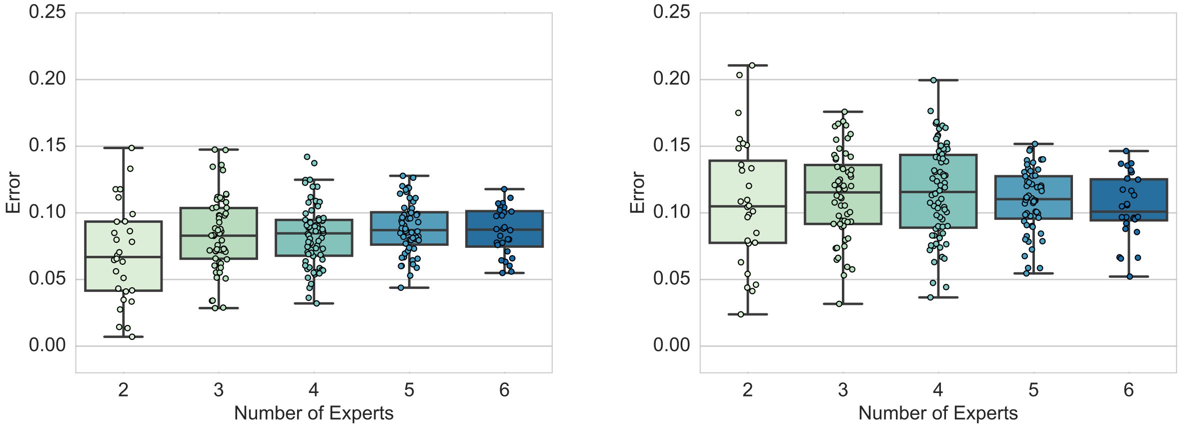


Figure 17. Absolute errors of MLE *(left) and* AGR *(right) for all possible subsets of between two and six experts, on the Social Media II dataset.*

5.4 FAKE NEWS

Dataset For our last dataset, we used the Buzzfeed Fact Check dataset[2]. The dataset consists of posts from a variety of mainstream, left-wing, and right-wing Facebook pages. The dataset was labeled by Buzzfeed staff into four classes, “Mostly True”, “Mixture of True and False,” “Mostly False,” and “No Factual Information.” Labels were determined by Buzzfeed staff and required fact-checking and providing citations for claims made in the article.[3] To simplify the problem for our experts (who could not be expected to fact-check a sample of 100 or 200 articles), we combined the “Mostly False” and the “Mixture of True and False” classes into a single “Fake” class, relabeled “Mostly True” as “Not Fake,” and excluded the “No Factual Content” class. The descriptions of each class are given in Table 4.

Classifier and experts We used a support vector machine as our classifier, trained on TF-IDF features from the document title and body as well as other counts (number of authors, number of capitalized words, etc.) and the number of Facebook comments, likes, and shares. For this

[2] `https://github.com/BuzzFeedNews/2016-10-facebook-fact-check`

[3] We consider this a mostly reliable procedure, though there is obviously some subjectivity in determing between “Mostly True” and “Mixture of True and False” as well as “Mixture of True and False” and “Mostly False.”

TABLE 4

Description of the classes in the original Buzzfeed dataset and which classes we used in our “Fake” and “Not Fake” classes.

Our Class	Buzzfeed Class	Description
Fake	Mostly False	“Most or all of the information in the post or in the link being shared is inaccurate. This should also be used when the central claim being made is false.”
Fake	Mixture of True and False	“Some elements of the information are factually accurate, but some elements or claims are not. This rating should be used when speculation or unfounded claims are mixed with real events, numbers, quotes, etc., or when the headline of the link being shared makes a false claim but the text of the story is largely accurate. It should also only be used when the unsupported or false information is roughly equal to the accurate information in the post or link. Finally, use this rating for news articles that are based on unconfirmed information.”
Not Fake	Mostly True	“The post and any related link or image are based on factual information and portray it accurately. This lets them interpret the event/info in their own way, so long as they do not misrepresent events, numbers, quotes, reactions, etc., or make information up. This rating does not allow for unsupported speculation or claims.”
N/A	No Factual Content	“This rating is used for posts that are pure opinion, comics, satire, or any other posts that do not make a factual claim. This is also the category to use for posts that are of the Like this if you think... variety.”

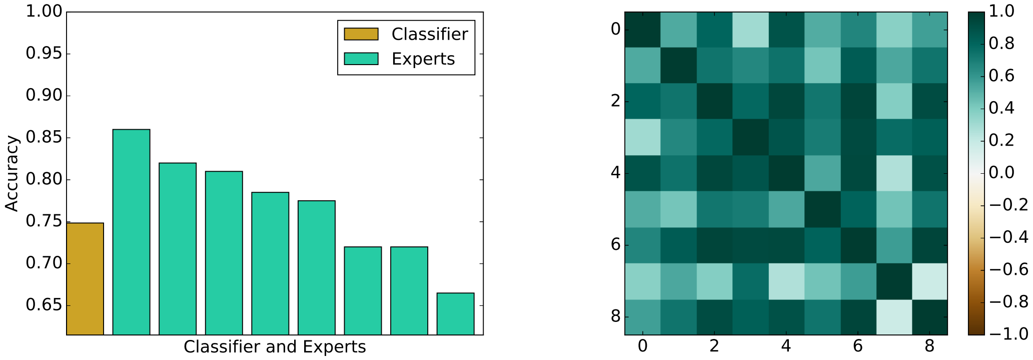


Figure 18. Left: Accuracy of the classifier on the entire test dataset and accuracies of each expert on the sample from the test dataset. Right: A class-averaged version of Yule's Q statistic between all pairs of classifier/experts.

experiment, we collected 200 labels from eight experts. As before, experts were given the raw text, rather than the features described above. We also provided the experts with the BuzzFeed class descriptions, to prevent experts from developing differing interpretations of what "Fake News" is. On the left of Figure 18 we show the accuracy of the classifier and each expert. The classifier is about 75% accurate while the experts are from 67–86% accurate. Unlike our other experiments, the classifier does not outperform the experts. On the right, we show the degree of conditional dependence between classifier/experts. Again, we observe strong positive dependence between classifier/experts. Like the two social media experiments, there does not seem to be any structure.

Estimator errors We show the performance of the estimators on 10,000 sets of resamples (of size 200) from the dataset in Figure 19. We quickly notice that each estimator consistently overestimates the true value, unlike in the other experiments, where we mostly underestimated the true value. This is perhaps explained by the classifier not out-performing the experts. The positive bias may also be partly explained by the sample, as EMP is positively biased. The estimators all have median errors of less than 0.10 and never make an error larger than 0.20.

We show the mean absolute error of each estimator in Table 5. MLE and MV perform best on this dataset, with mean absolute errors of 0.046. COV is the worst, with mean absolute error of 0.093.

Increasing the number of experts We apply each estimator to every possible subset of experts of size $2, 3, \ldots, E-2$. In Figure 20, we show boxplots of the results for MLE and AGR. For MLE, the median error remains approximately constant as the number of experts increases, while the variance declines. For AGR, we also see roughly constant median error. The variance appears larger for even numbers of experts than for odd numbers of experts.

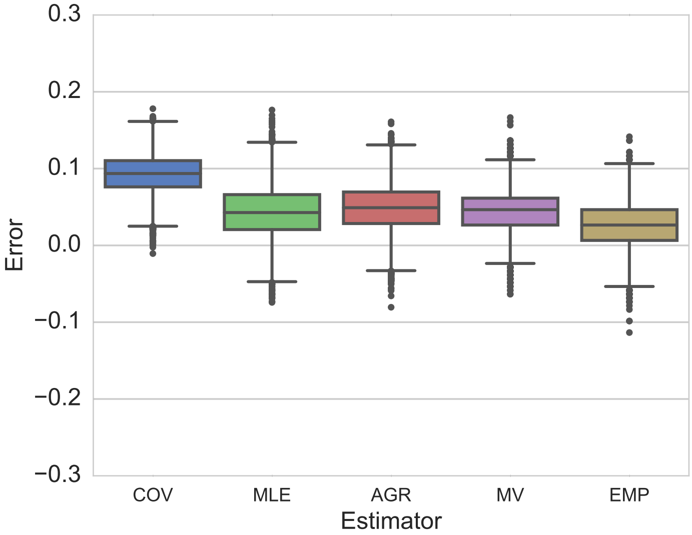


Figure 19. Signed errors for 10,000 resamples of the Fake News dataset, for each estimator.

TABLE 5

Mean absolute errors of the estimators across 10,000 resamples of the Fake News dataset.

Estimator	Mean Absolute Error
COV	0.093
MLE	0.046
AGR	0.050
MV	0.046

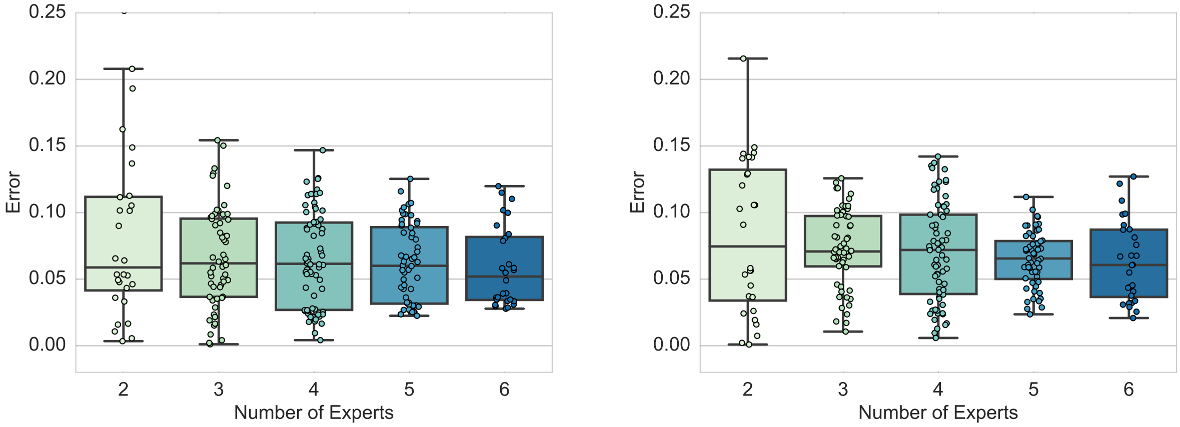


Figure 20. Absolute errors of MLE *(left) and* AGR *(right) for all possible subsets of between two and six experts, on the Fake News dataset.*

5.5 SUMMARY

The simplest possible analysis of these estimators on our datasets is to compute the error for each estimator using all samples. We show this in Table 6. We include AGR-OPT and BEE on Fake News, since it is a binary problem. To summarize, MLE performs best on three of the four datasets and COV is best on the other. MLE is always better than MV, while COV and AGR are better except on Fake News. Because of the small sample size, we should not read too much into the performance of the estimators based on these results. It is clear from the resampling approach taken in the preceding sections that each estimator has significant variability. Nevertheless, we present the results for completeness.

We end our experimental section by drawing the following broad conclusions about the estimators and experts:

1. The estimators are not a replacement for ground truth. Except for Social Media I, the estimators have noticeably larger errors than EMP.

2. The median performance of MLE is often one of the best, but it tends to have a larger variance. In our final comparison, MLE has the smallest error in three of the four experiments.

3. The mean absolute error over resamples of the datasets for COV, MLE, and AGR is sometimes, but not always, better than MV. MLE is better on three of the four datasets (and essentially ties on the other) while AGR and COV are better on two of the four. We argue that the fact that these estimators do not always outperform MV is not surprising. It is well known in the ensemble classifier literature that MV often performs very well at inferring the true label. In addition, we saw in our simulations that the biggest difference between MV and the other estimators occurs when experts have low accuracy and the classifier has high accuracy. Since the experts and classifier usually had reasonably similar performance, we are not surprised that MV performs about as well as the other estimators.

4. Human experts almost always exhibit strongly positive conditional dependence. Thus, it seems clear that assuming experts are conditionally independent is not a valid assumption.

5. AGR-OPT and BEE have performance similar to the other estimators on Fake News (for the simple comparison in Table 6, but since they only apply to binary problems, they are of limited interest to us.

6. Increasing the number of experts does not have as clear of an effect as we would have anticipated. It sometimes helps, but also sometimes appears to make little practical difference.

TABLE 6

Errors for estimating the accuracy of the classifier, using all samples. The smallest error for each experiment is in bold.

	COV	MLE	MV	AGR	AGR-OPT	BEE
Job Category	−0.077	**−0.055**	−0.112	−0.089	-	-
Social Media I	0.027	**0.016**	−0.044	−0.033	-	-
Social Media II	**−0.062**	−0.088	−0.108	−0.083	-	-
Fake News	0.093	**0.042**	0.047	0.049	0.098	0.050

6. DIAGNOSTIC TESTS

In this section, we discuss whether certain diagnostics (inter-rater reliability and an algorithm to detect dependent groups of experts) can help calibrate a practitioner's expectations about how well the estimators will perform.

6.1 INTER-RATER RELIABILITY

We briefly describe the usefulness (or lack thereof) of inter-rater reliability statistics in the context of the estimators considered in this paper. Traditionally, a study involving expert judgment may not be considered trustworthy if experts do not attain a high inter-rater reliability score. In Table 7, we compute Fleiss' κ for the experts in our four experiments. We also show the results for data simulated from the copy model and the group model, for increasing amounts of dependence (increasing the correlation parameter ρ in the copy model and decreasing the number of groups in the groups of experts model).

In the experiments, the experts have moderate inter-rater agreement. However, we strongly caution making any interpretation of these numbers. To explain why, observe that in the simulated data, as we *increase* the parameter that controls the amount of dependence in the simulation, while holding accuracy constant, inter-rater reliability also *increases*. We argue that this is not surprising. By applying one of these estimators, one is assuming that experts are not particularly accurate (if experts are very accurate, then good estimation of accuracy can be obtained simply by comparing the classifier labels to one set of expert labels). As a result, we expect disagreement between experts and correspondingly low values of inter-rater reliability. In fact, if experts are both inaccurate *and* exhibit high inter-rater reliability, they are likely making similar mistakes, meaning that they are not conditionally independent of one another, which we have observed has a negative impact on the effectiveness of the estimators. Thus, if one observes high inter-rater reliability, this can in fact indicate that the experts are both inaccurate and dependent.

6.2 DETECTING DEPENDENCE BETWEEN EXPERTS

Recall that in Section 4, we saw that in a variety of simulation models, conditional dependence between experts increases error. In the experiments contained in Section 5, we saw that actual human experts do not satisfy the conditional independence assumption inherent in most of the estimators. It is of interest then, to consider whether it is possible to detect conditional dependence between experts (without using the true labels). In what follows, we discuss an unsupervised algorithm to detect groups of dependent experts and whether the presence of groups of dependent experts is helpful to the estimators.

6.2.1 Detecting groups of experts

The only direct attempt to detect a dependence model of which we are aware is an algorithm by Jaffe et al. [18]. The authors postulate that a collection of classifiers may naturally consist of several distinct "groups" (we can think of this as experts that primarily use a similar group of features in their decision making). Their algorithm for detecting groups of experts is highly related to the COV estimator.

Method First, recall the matrix S defined in Section 3.4 that has off-diagonal entries equal to those of the classifier/expert covariance matrix. Jaffe et al. [18] show the rather technical result

$$\mathsf{det}\left(\begin{bmatrix} S_{ij} & S_{il} \\ S_{kj} & S_{kl} \end{bmatrix}\right) = S_{ij}S_{kl} - S_{kj}S_{il}$$

is zero if and only if

- at least three of the classifiers belong to the same group,
- or if $g(i) \neq g(j)$, $g(j) \neq g(k)$, $g(k) \neq g(l)$, and $g(l) \neq g(i)$.[4]

With this relationship in mind, if we consider

$$D_{ij} = \sum_{k,l} |S_{ij}S_{kl} - S_{kj}S_{il}|,$$

we can see that when $g(i) = g(j)$, then $S_{ij}S_{kl} - S_{kj}S_{il}$ can be zero only when three of the classifiers are in the same group. On the other hand, when $g(i) \neq g(j)$, then $S_{ij}S_{kl} - S_{kj}S_{il}$ can be zero when either of the two above conditions hold. Thus, we expect d_{ij} to be large when $g(i) = g(j)$ and small if $g(i) \neq g(j)$. Jaffe et al. [18] bound the size of d_{ij} below when $g(i) = g(j)$ and from above when $g(i) \neq g(j)$, for a specific group structure. Once the matrix D is computed, the authors apply a clustering algorithm to infer the group structure.

[4] We use $g(i)$ to denote the group that classifier/expert i belongs to.

TABLE 7

Fleiss' κ for experts from the experiments, as well as simulated experts from the Copy and Group models.

Experiment	Fleiss' κ	Simulation	Fleiss' κ
Job Category	0.34	Copy ($\rho = 0.0$)	0.38
Soc. Media I	0.61	Copy ($\rho = 0.25$)	0.49
Soc. Media II	0.62	Copy ($\rho = 0.50$)	0.54
Fake News	0.47	Copy ($\rho = 0.75$)	0.73

Simulation	Fleiss' κ
Group (4 groups)	0.27
Group (3 groups)	0.30
Group (2 groups)	0.35
Group (1 group)	0.49

TABLE 8

Proportion of times that spectral clustering returns the correct group structure for increasing numbers (k) of groups.

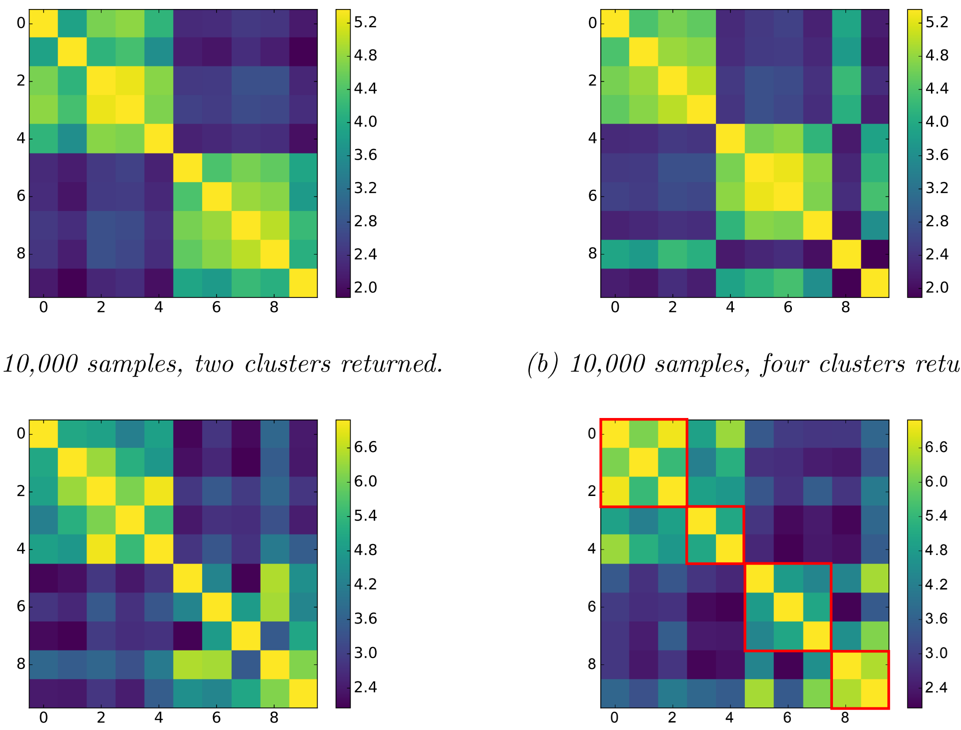
k	2	3	4	5	6	7	8	9
Proportion correct	0.999	0.992	0.973	0.975	0.791	0.687	0.573	0.647

Simulations To see whether this idea works in practice, we consider the following simple simulation. Consider a problem with 10 experts that each label 200 data points. The experts are evenly distributed amongst k groups. (So when $k = 4$, there are two groups of three and two groups of two.)

For each $k = 2, \ldots, 9$, we drew 1000 sets of 200 samples. For each set, we compute the matrix D and apply spectral clustering (using the correct number of clusters) to determine groups. Table 8 shows the percent correct for $k = 2, \ldots, 9$. It appears that, despite the small sample size, the algorithm recovers the exact group assignment with high probability up until $k = 6$. This is encouraging, as it implies that with a human-in-the-loop to examine the output of the clustering, it *may* be possible to recover the assignment function even if the true number of clusters is not known. In practice however, making this choice may be difficult, especially when there are only a small number of samples. In Figure 21, we show the groups returned with spectral clustering when $k = 2$ and the number of groups requested is 2 and 4 with either 10000 or 200 samples. It is visually evident (for 10000 samples) that the first clustering is superior to the second, but somewhat less clear when the number of samples is 200.

Job category experiment In Figure 22, we show the clustering results on the matrix D for the job category experiment for two, three, and four groups. It was for this experiment that we previously noted a pattern in the dependency structure. Visually, four groups clearly seems to be a poor fit, and it appears that two groups is a somewhat better fit than three groups. We consider the "right" answer to be two groups. To see why, consider Figure 23, which shows the results of clustering on the matrix of pairwise Yule's Q statistic with two, three, and four groups. It is abundantly clear that two is the best fit.

Assuming we can identify the presence of groups, the natural question is, does having multiple groups of experts help or hurt our estimate? To answer the question, we compare three scenarios using 100 resamples of the Job Category data. First, estimate classifier accuracy using the classifier and three experts from the classifier's group. Second, estimate classifier accuracy using the classifier and all three experts from the other group. Third, estimate classifier accuracy using the classifier, one expert from the classifier group and two experts from the other group. The results for MLE and AGR are shown in Figure 24. It is encouraging to see that for both estimators, and in particular AGR, the performance is much better when experts come from separate groups, rather than one or the other. The result is intuitive, as the experts in one group are conditionally independent given



(a) 10,000 samples, two clusters returned.

(b) 10,000 samples, four clusters returned.

(c) 200 samples, two clusters returned.

(d) 200 samples, four clusters returned.

Figure 21. Clusters returned by spectral clustering (highlighted in red for clarity in the bottom right) for 10,000 and 200 samples, when two and four clusters are requested. The correct number of clusters is two.

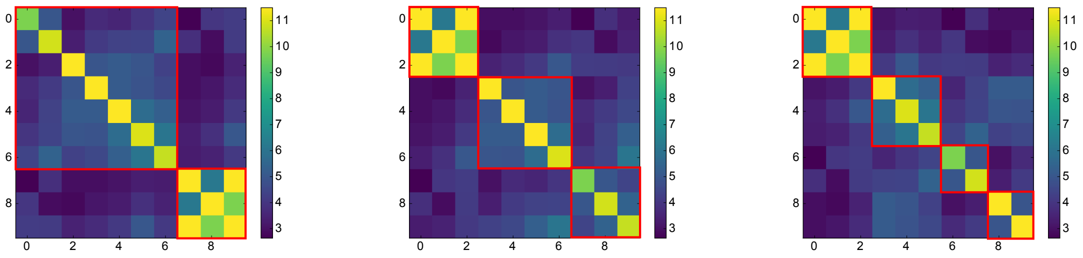


Figure 22. Clusters returned by spectral clustering (highlighted in red for clarity), when two, three, and four clusters are requested, using the matrix D for the job category experiment.

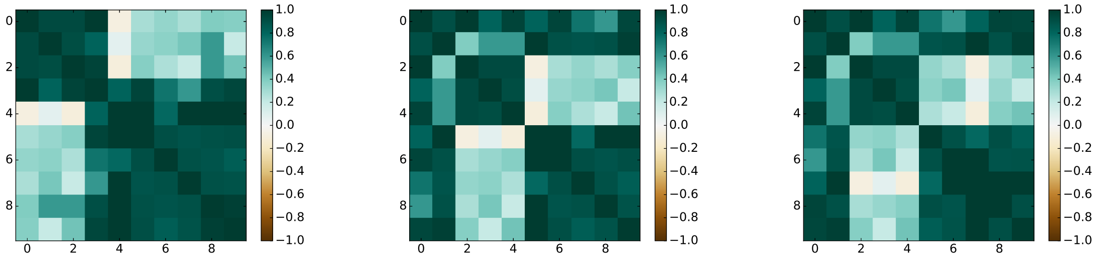


Figure 23. Clusters returned by spectral clustering, when two, three, and four clusters are requested, using the matrix of pairwise Yule's Q statistic for the job category experiment.

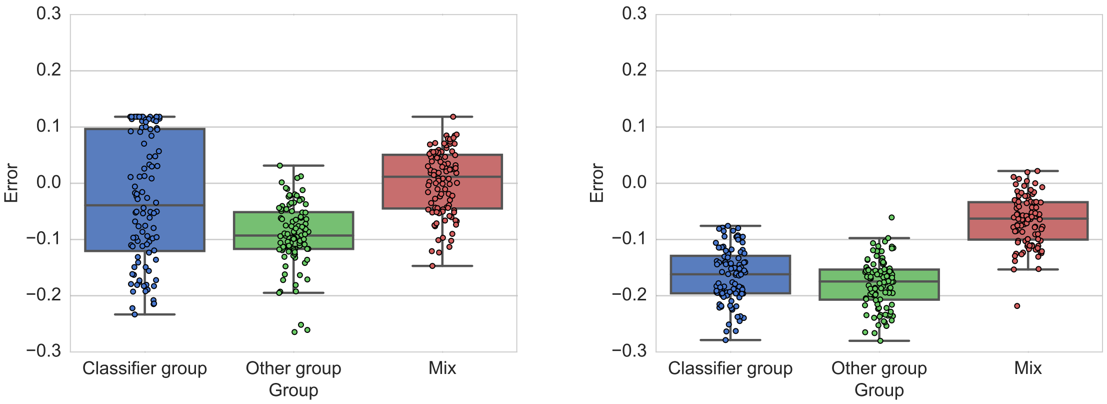


Figure 24. Left: Error of MLE when applied to experts that come from only the classifier group, only the other group, or to experts from both groups. Right: Error of AGR for the same experiment.

the true label from the experts in the other group, which is a desirable property for all estimators. Results for the other estimators are similar to those described here.

Unfortunately, for our other datasets, it does not appear that such a group structure exists. As a result, our investigation into this phenomena on real data is limited to this one example.

6.2.2 Feature partitioning to induce conditional independence

Given the observation from the previous section that experts coming from different groups is helpful to the estimators, we consider whether it is possible to structure an experiment with human experts to encourage the formation of multiple groups. In related work, authors have suggested attempting to induce conditional independence between automated classifiers by training them on different feature sets. For instance, in [21], the authors train separate classifiers on the title and body of text documents. In [22], the author introduces a greedy algorithm to identify subsets of features such that each subset shares information with the true label, but shares as little information as possible with the other feature subsets. For the estimators described in Section 3, we are interested in the case where human experts provide labels. In the machine learning literature at least, it appears that most tasks given to experts (or crowd workers) come from the image or text domain; see for example [9, 19, 23, 24].[5] While automated machine learning algorithms may use a variety of types of features to solve these problems, it is difficult to imagine providing humans with anything other than the raw image or text. Thus, it does not seem that dividing features into groups is an extensible approach for most classification problems involving human experts.

[5] There are certainly exceptions to this, for example the Job Category classifier we used in Section 5.

7. CONCLUSION AND FUTURE WORK

Based on the simulations and experiment performed in this paper, we draw the following broad conclusions. For clarity, we organize these into three areas:

Conditional dependence

1. Conditional dependence between experts clearly impacts the estimators. In each of our three simulations, we saw larger errors as experts became more dependent.
2. In our experiments, we almost always saw strong conditional dependence between experts, indicating that real human experts cannot be assumed to be conditionally independent.
3. Ideally, we would want to control this dependence in the design of experiments. The only work of which we are aware that attempts to address this issue suggests trying to encourage independence by providing different experts with different features. We do not believe this is a reasonable approach as most problem for which humans can easily provide labels involve either text or images, which are not amenable to feature partitioning.
4. Inter-rater reliability should not be used to calibrate expectations about the effectiveness of the estimators, because it increases as experts become more dependent.
5. If expert dependence is according to the groups of experts model, it appears that this can be detected in some cases by the approach described in Section 6. We observed experts that appeared to correspond to this model in the job category experiment, but not in any of the others.

Expert and classifier performance

1. All estimators have larger errors when the experts have low accuracy. However, there is not a consistent pattern in terms of whether this manifests as overestimates or underestimates of accuracy. We observed both in the simulations.
2. The above observation is somewhat disappointing, if not surprising, as it implies that it is not possible, in general, to achieve a small error if the classifier is significantly more accurate than the experts.

Estimators

1. In our simulations, the remaining estimators, including MV, perform reasonably well, particularly for the groups of experts model and the difficulty model, as long as experts are at least close to the accuracy of the classifier. COV, MLE, and AGR usually outperform MV. The difference is greater when experts are less accurate than the classifier.

2. In our experiments, the estimators all perform perhaps surprisingly well. In particular, errors (on the full dataset) for MLE, COV, and AGR are always less than 0.10 and often less than 0.05.

3. The experiments also support the use of COV, MLE, or AGR over MV, as they outperformed, in terms of mean absolute error, MV on two (COV and AGR) or three (MLE) of our experiments.

4. There does not seem to be strong experimental evidence to support using one of COV, MLE, or AGR over the others. We note that MLE tends to have the best mean and median performance but in some cases has a large variance.

7.1 FUTURE WORK

There are many opportunities for future work in this area. One obvious extension is to consider estimating class-conditional accuracy. Several of the estimators, COV, MLE, and BEE, can easily be extended to do this. We speculate that considerably more samples would be required to reliably estimate each of the class-conditional accuracies. Another interesting question is whether a clear relationship between the bias of the estimators and the difference in accuracy between the classifier and experts exists. Finally, given that we can in some cases detect groups of experts in an unsupervised manner, it would be interesting to explore whether the method for doing so can be extended or has relevance to other models of dependence.

REFERENCES

[1] A.P. Dawid and A.M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1), 20–28 (1979).

[2] P. Donmez, G. Lebanon, and K. Balasubramanian, "Unsupervised supervised learning I: Estimating classification and regression errors without labels," *J. Mach. Learn. Res.* 11, 1323–1351 (2010).

[3] E.A. Platanios, A. Blum, and T.M. Mitchell, "Estimating Accuracy from Unlabeled Data," in *Conference on Uncertainty in Artificial Intelligence* (2014).

[4] A. Jaffe, B. Nadler, and Y. Kluger, "Estimating the accuracies of multiple classifiers without labeled data," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* (2015).

[5] P.E. Lehner, "Estimating the accuracy of automated classification systems using only expert ratings that are less accurate than the system," *Journal of Modern Applied Statistical Methods* 14(1) (2015).

[6] E.A. Platanios, A. Dubey, and T.M. Mitchell, "Estimating accuracy from unlabeled data: A bayesian approach," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (2016), pp. 1416–1425.

[7] G.J. Ciach and W.F. Krajewski, "On the estimation of radar rainfall error variance," *Advances in Water Resources* 22(6), 585–595 (1999).

[8] N.E. Bowler, "Explicitly accounting for observation error in categorical verification of forecasts," *Monthly Weather Review* 134(6), 1600–1606 (2006).

[9] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (2010), pp. 25–32.

[10] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *SIGIR 2011 Workshop on Corwdsourcing for Information Retrieval* (2011).

[11] H.C. Kim and Z. Ghahramani, "Bayesian classifier combination," in *Proceedings of the 15th International Confernce on Artificial Intelligence and Statistics* (2012).

[12] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web* (2014), pp. 155–164.

[13] W. Bi, L. Wang, J.T. Kwok, and Z. Tu, "Learning to predict from crowdsourced data," in *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence* (2014).

[14] P.G. Moreno, A. Artés-Rodríguez, Y.W. Teh, and F. Perez-Cruz, "Bayesian nonparametric crowdsourcing," *J. Mach. Learn. Res.* 16(1), 1607–1627 (2015).

[15] M. Venanzi, J. Guiver, P. Kohli, and N.R. Jennings, "Time sensitive-bayesian information aggregation for crowdsourcing systems." *Journal of Artificial Intelligence Research* 56, 517–545 (2016).

[16] U. Shaham, X. Cheng, O. Dror, A. Jaffe, B. Nadler, J. Chang, and Y. Kluger, "A deep learning approach to unsupervised ensemble learning." in *Proceedings of the 33rd International Conference on Machine Learning* (2016).

[17] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, "Ranking and combining multiple predictors without labeled data," *Proceedings of the National Academy of Sciences* 111(4), 1253–1258 (2014).

[18] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised ensemble learning with dependent classifiers," in *Proceedings of the 19th International Confernce on Artificial Intelligence and Statistics* (2016).

[19] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (2009), pp. 2035–2043.

[20] D. Aldous, "Exchangeability and related topics." in *Ecole d'Ete St Flour*, Springer (1985).

[21] P. Bommannavar, A. Kolcz, and A. Rajaraman, "Recall estimation for rare topic retrieval from large corpuses," in *2014 IEEE International Conference on Big Data* (2014).

[22] S.V. Mane, *False negative estimation: theory, techniques and applications*, Ph.D. thesis, UMN (2008).

[23] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems* (2010), pp. 2424–2432.

[24] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2008), pp. 254–263.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 31-01-2018	**2. REPORT TYPE** Technical Report	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE

Estimating Classifier Accuracy Using Noisy Expert Labels

5a. CONTRACT NUMBER
FA8721-05-C-0002 and/or FA8702-15-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

J.T. Holodnak
J.T. Matterer
W.W. Streilein

5d. PROJECT NUMBER
2716

5e. TASK NUMBER
272

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421-6426

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-1225

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Intelligence Advanced Research Projects Activity
Office of the Director of National Intelligence
Washington DC 20511.

10. SPONSOR/MONITOR'S ACRONYM(S)
IARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

13. ABSTRACT

In this work, we present an empirical comparison of statistical methods that estimate the accuracy of a classifier using noisy expert labels. We are motivated by the application of machine learning to difficult problems for which even experts may be unable to provide an authoritative label for every data instance. Several estimators have been recently proposed in the literature, but prior empirical work to evaluate the applicability of these estimators to real-world problems is limited. We apply the estimators to labels simulated from three models of the expert labeling process and also four real datasets labeled by human experts. Our simulations reveal the importance of the accuracy of the classifier relative to the experts and confirm that conditional dependence between experts negatively impacts estimator performance. On two of the real datasets, the estimators clearly outperformed the baseline majority vote estimator, supporting their use in applications. We also briefly examine the utility, in terms of increasing or decreasing confidence in an estimator's output, of a few diagnostics that can be applied to the expert labels.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unclassified	64	**19b. TELEPHONE NUMBER** *(include area code)*

This page intentionally left blank.